Enriching Unsupervised User Embedding via Medical Concepts

Xiaolei Huang   XIAOLEI.HUANG@MEMPHIS.EDU
University of Memphis, United States

Franck Dernoncourt   DERNONCO@ADOBE.COM
Adobe Research, United States

Mark Dredze   MDREDZE@CS.JHU.EDU
Johns Hopkins University, United States

Abstract

Clinical notes in Electronic Health Records (EHR) present rich documented information of patients to inference phenotype for disease diagnosis and study patient characteristics for cohort selection. Unsupervised user embedding aims to encode patients into fixed-length vectors without human supervisions. Medical concepts extracted from the clinical notes contain rich connections between patients and their clinical categories. However, existing *unsupervised* approaches of user embeddings from clinical notes do not explicitly incorporate medical concepts. In this study, we propose a concept-aware unsupervised user embedding that jointly leverages text documents and medical concepts from two clinical corpora, MIMIC-III and Diabetes. We evaluate user embeddings on both extrinsic and intrinsic tasks, including phenotype classification, in-hospital mortality prediction, patient retrieval, and patient relatedness. Experiments on the two clinical corpora show our approach exceeds unsupervised baselines, and incorporating medical concepts can significantly improve the baseline performance.

Data and Code Availability In this study, we experiment with two clinical datasets that are publicly available, Medical Information Mart for Intensive Care (MIMIC)-III (Johnson et al., 2016) and Clinical Trial Cohort Selection (Diabetes) (Stubbs et al., 2019).[1] The data description and processing details is in Section 2. Our experimental codes, supplementary analyses, and data access instructions are at https://github.com/xiaoleihuang/UserEmb_Explainable.

---

1. We denote the cohort selection dataset as Diabetes because patients share the same syndrome, diabetes. Diabetes data link: https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/, MIMIC-III data link: https://physionet.org/content/mimiciii/1.4/.

1. Introduction

User embedding aims to learn fix-length vectors for users and maps all user information into a unified vector space (Pan and Ding, 2019; Li and Zhao, 2020; Si and Roberts, 2020). Patient representations inferred by user embeddings have shown broad applications in medical diagnosis (Choi et al., 2016, 2017), hospital stays (Wang et al., 2019), and patient readmission (Darabi et al., 2020). The supervised learning approaches require a large amount of annotated datasets, yet obtaining human annotations for clinical notes can be labor-intensive, time-consuming, and expensive. Additionally, International Classification of Diseases (ICD) codes or other human clinical annotations have annotating issues including noisy (Harutyunyan et al., 2019), error-prone (Birman-Deych et al., 2005), and low inter-rater agreement (Stubbs et al., 2019). Quality of data annotations can impact the robustness of supervised embedding models.

Unsupervised user embedding learns patient representations from clinical notes without human supervision. The embedding aims to capture semantic relations between patients and their clinical records. To represent clinical notes, models derive features from topic model (Blei et al., 2003), paragraph vectors (Le and Mikolov, 2014), word2vec embeddings (Mikolov et al., 2013). Finding hidden structures to distinguish patients is a challenge for unsupervised user embeddings to compress patient features from clinical notes. Unsupervised approaches (Sushil et al., 2018; Lei et al., 2018) utilize autoencoders by feature reconstruction, while ignoring other structures like medical concepts.

Medical *concepts*, which consist of one or more tokens, are basic description units for medical information, such as disease symptom and clinical drug. The concepts contain rich information of medical ontology that owns strong connections with phenotype infer-

ence. Medical concept embeddings treat concepts as special tokens or phrases and encode them as word vectors. The representations are effective in downstream tasks of health diagnosis (Choi et al., 2016, 2020) and readmission forecast (Choi et al., 2017; Zhang et al., 2020). However, unsupervised user embeddings have not explicitly consider medical concepts, which can lose semantic information of clinical conditions.

In this study, we propose a *concept-aware user embedding* (CAUE) model that jointly handles clinical notes and medical concepts under the multitask learning framework. Our unsupervised method applies contrastive learning (Logeswaran and Lee, 2018; Chen et al., 2020) and negative sampling to leverages two levels of information sources, patient-document and patient-concept. Patient-document task encodes sequential information of patients to guide models recognizing patients of clinical notes, and patient-concept task treats medical concepts as a medium point to constrain patients. To model unusually lengthy clinical documents, we propose a data augmentation approach, random split, to cut each note into random sizes of snippets. We evaluate on multiple clinical datasets with two extrinsic tasks, phenotype and mortality predictions, and two intrinsic tasks, patient relatedness and retrieval. Research suggests that intrinsic evaluations of embedding models can reduce parameter biases of extrinsic evaluations for controlling fewer hyperparameters (Schnabel et al., 2015). The four evaluations show that our proposed approach achieves the best performance overall and highlight the effectiveness of medical concepts. Our ablation analysis shows that integrating medical concepts can generally improve effectiveness of baseline models.

2. Data

In this section, we introduce the two clinical data (Diabetes and MIMIC-III) and report data processing steps. We obtain the Diabetes data (Stubbs et al., 2019) from Track 1 of the 2018 National NLP Clinical Challenges (n2c2) shared task, which contains a collection of longitudinal patient records, with 2 to 5 text documents per patient. The clinical documents summarize diagnosis results and describe clinical conditions indicating if patients have coronary artery disease. Each patient in the Diabetes data has 13 selection criteria annotations, which follow ClinicalTrials.gov and relate to diabetes and heart disease. MIMIC-III is a relational database that collects a large set of intensive care unit (ICU) patients from the Beth

Israel Deaconess Medical Center between 2001 and 2012.[2] The data includes various patient information such as demographics, laboratory results, mortality, and clinical notes. Clinical notes are unstructured narrative documents describing patients' ICU stays, including radiology and discharge summaries. Like Diabetes, each MIMIC-III document has a list of medical annotations, International Classification of Disease, 9th Edition (ICD-9). The ICD-9 annotations indicate symptoms and disease types of patients during their ICU stays. Both data corpora have been de-identified following Health Insurance Portability and Accountability Act (HIPAA) standards, and any dates were time-shifted by a random amount.

We use MetaMap (Aronson and Lang, 2010) to extract medical concepts by the Unified Medical Language System Metathesaurus (UMLS) (Bodenreider, 2004).[3] MetaMap extracts medical concepts with their Concepts Unique Identifier (CUI) linked to the UMLS, confidence scores, and corresponding medical types. To improve accuracy and reduce the ambiguity of extracted concepts, we enable the word sense disambiguation (WSD) module. The WSD module checks if concepts are semantically consistent with surrounding texts and automatically removes ambiguous entities. Each concept may have one or more tokens and associate with at least one concept type. We lowercase all extracted medical concepts and empirically removed non-symptom-related entities, including digits-only sequences, language, and temporal concepts.[4]

In this study, we preprocess medical annotations and clinical notes for both corpora. For MIMIC-III, we follow previous work (Mullenbach et al., 2018) to keep discharge summary, which merges enough details of a single admission for each patient. ICD-9 labels for the MIMIC-III have over 15K different codes, which are highly dimensional and sparse. To reduce sparsity and noise, we follow previous work (Harutyunyan et al., 2019) to merge ICD-9 codes according to the Health Cost and Utilization (HCUP) Clinical Classification Sofware (CCS).[5] The predefined HCUP-CCS schema

---

2. We access the data at https://physionet.org/content/mimiciii/1.4/. MIMIC has different versions, and we experiment with the stable version, MIMIC-III 1.4. Its release year is 2016.

3. While we have alternative toolkits, such as cTAKES (Savova et al., 2010) and CLAMP toolkit (Soysal et al., 2017), however, they did not adapt to the new authentication system of UMLS when this study started.

4. MetaMap concept types: https://metamap.nlm.nih.gov/Docs/SemanticTypes_2018AB.txt

5. https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp

Dataset	Document Statistics			User Stats				Concept Stats	
	Doc	Vocab	Token-stats	User	Age	F	U-label	Concept	Type
Diabetes	1265	34592	2426, 483, 42	288	63.13	0.45	10	68938	89
MIMIC-III	54888	390237	7522, 1263, 50	48807	62.47	0.44	276	10761211	94

Table 1: Data summary of Diabetes and MIMIC-III by documents, patients, and concepts. The "Token-stats" includes maximum, median and minimum lengths of processed clinical notes. We present values of patient count (User), averaged age (Age), gender ratio of female (F) and male (M), number of unique labels (U-label), number of concepts (Concept), and number of unique concept types (Type).

clusters ICD-9 procedure, billing, and diagnostic codes into mutually exclusive, largely homogeneous disease categories. We follow previous work (Harutyunyan et al., 2019) to drop any patients who are younger than 18 years old. Two domain experts annotated each patient from the Diabetes data by 13 different clinical criteria. We empirically drop three of them because their inter-rater agreements are fewer than 0.5.[6] We apply the same preprocessing steps for clinical notes. Each patient associates with one or more notes and their medical annotations. The temporal shift is an attribute of the in-hospital records, MIMIC-III. To reduce temporal impacts and noises, we isolate a patient's records per visit as a separate patient. For each note, we lowercase all tokens, tokenize sequences by NLTK (Bird and Loper, 2004), remove person and hospital names, replace numbers and dates with placeholders ([NUM] and [DATE]), and remove repeated punctuation. We empirically drop any notes with less than 40 tokens.

We summarize data statistics in Table 1. MIMIC-III shows more varied statistics of clinical notes than the Diabetes dataset. For example, the range of token counts for MIMIC-III data is broader than Diabetes (50 to 7522 vs. 42 to 2426), and MIMIC-III has a larger vocabulary. Both datasets have approximately similar demographic distributions of age and gender. MIMIC-III data has more complex clinical annotations, while Diabetes data has a maximum of 10 labels for each patient. Comparing to health data from social media, such as Twitter, the clinical corpora have more senior user groups and more variations in documents, patients, and annotation information. For example, while the Twitter suicide users (Amir et al., 2017) only has binary labels, MIMIC-III data has 276 unique phenotypes.

---

6. The agreement rate of ENGLISH, MAKES-DECISIONS, and KETO-1YR are 0.46, 0.31, -0.1, respectively. Note that the main reason that causes a negative agreement score of KETO-1YR is the skewed distributions of annotations.

2.1. Privacy and Ethical Considerations

We do not release any clinical data associate with patient identities due to privacy and ethical considerations. Instead, we have released our code and provided detailed instructions to replicate our analysis and experiments. To protect user privacy, we have followed corresponding data agreements to ensure the proper data usage and experimented with de-identified data. Our experiments do not store any patient data and only use available text documents for research demonstrations. Except, our experiments use anonymized patient ID and clinical notes for training and evaluating user embeddings.

3. Concept-Aware User Embedding ($CAUE$)

In this section, we propose an **unsupervised** concept-aware user embedding ($CAUE$) that jointly models clinical notes and medical concepts under a multi-task learning framework. Methods (Si et al., 2020) to train unsupervised user embedding from clinical notes primarily focus on bag-of-words features without explicitly capturing semantic information of medical concepts. (Steinberg et al., 2021) utilizes medical codes of clinical records as input features to derive patient representations sequentially, while our focus is to model sequential language features, clinical notes. (Dligach and Miller, 2018) obtains user embeddings by training a supervised classifier using concept embeddings. However, our approach jointly models document and concept embeddings in an unsupervised manner. We propose two joint tasks: modeling language sequential dependency by neural document encoders (task 1: patient-document) and leveraging connections between medical concepts and patients (task 2: patient-concept). We present the model architecture and learning process in Figure 1.
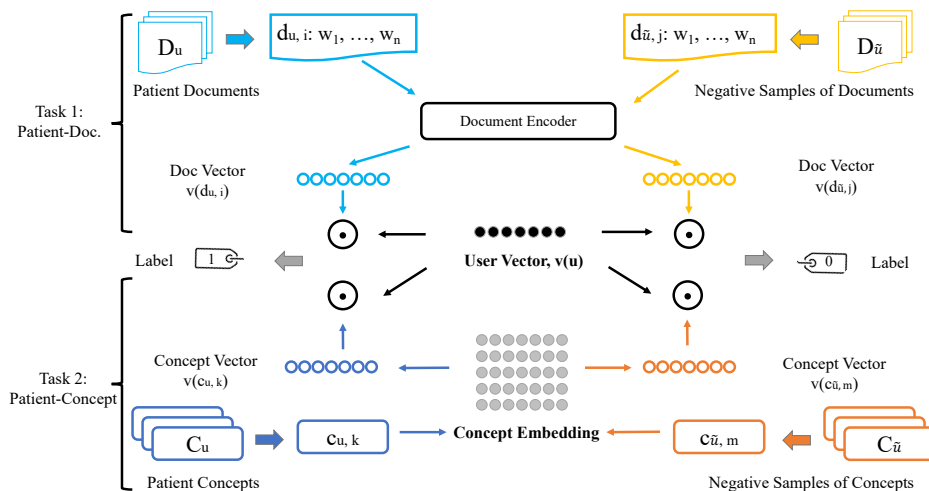
Figure 1: CAUE illustrations. All tasks center with the the patient, u. The arrows and their colors refer to the input directions and sources respectively. The $\odot$ is a similarity measurement.

3.1. Task 1: Patient-Document

The first task is to enforce models to understand contextual and sequential information of clinical notes so that models can summarize patients. Instead of training individual tokens, this task aims to model connections between users with documents. Given U be the list of patients, D_u be a list of clinical notes that describe the patient u, and we have $u \in U$, we can formalize our goal by maximizing the conditional probability of $P(D_u|u)$. Let $d_{u,i}$ as a clinical document (index as i) that describes the patient u clinical information, and each clinical note is independent with other notes in the collection D_u, our user-document task aims to maximize the following probability:

$$P(D_u|u) = \sum_{i=1}^{|D_u|} log P(d_{u,i}|u) \qquad (1)$$

To maximize the probability, we convert the task to a prediction task. Given a user, a model predicts a probabilistic distribution of a user vector over the patient's clinical notes so that the clinical note summarizes its target patient.

Considering the large size of the patients and clinical notes, we approximate our optimization objectives by the negative sampling and contrastive learning. The negative sampling is to randomly choose different documents from the existing corpus besides the target clinical document. The *contrastive learning* (Logeswaran and Lee, 2018; Chen et al., 2020)

is to predict if a sampled document or sentence is a context of the target. Previous study (Logeswaran and Lee, 2018) predicts the context sentence from several sampled or generated counterfactuals, which are stylistically similar to the context sentence but express different meanings. In this study, we treat patients as our target and clinical documents as context so that learning process will align patient representations toward their correct context, clinical documents. To generate counterfactuals, we choose two types of information: document and token. For the document level, we sample 3 snippets as counterfactuals by randomly choosing documents from the other patients. For the token level, we randomly replace tokens in the original document, where the sampling tokens are the vocabulary of each corpus. Instead of predicting each clinical document over all patients, the negative sampling enables model predictions over a few samples. This converts our prediction task into predicting if our model can identify if a text document describes the target patient or not. Let V as embeddings for documents, users and concepts, we can derive a document vector $v(d)$ for the document d_u, a user vector $v(u)$ for the patient u, and a concept vector $v(c)$ for the concept c. Then we can treat the task as a binary classification problem and minimize loss values by the binary cross-entropy:

$$\mathcal{L}(u, d) = -log(\sigma(v(u) \cdot v(d_u))) \\ - log(1 - \sigma(v(u) \cdot v(d_{\tilde{u}}))) \qquad (2)$$

66

where σ is a non-linear function and $d_{\tilde{u}}$ refers to noise documents. We use the sigmoid function as our non-linear function to normalize values of the dot production. The binary classification task will predict if a sampled clinical document describes correctly the target user, where one means correct match and 0 means incorrect descriptions.

We utilize a document encoder to derive document vectors $v(d)$ to learn language dependency across sequential tokens. Let a document d has a sequence of words, $d = \{w_1, w_2, ..., w_n\}$. Then we can derive document vectors by $v(d) = f_d(v(w_1), ..., v(w_n))$, where f_d is the document encoder and $v(w)$ is a word vector. Finally, our document encoder can learn sequential information of each clinical note into a unified vector, $v(d)$.

3.2. Long Text Modeling via Random Split

The clinical notes have a much longer document length (e.g, 7522 in our case) comparing to Twitter messages. Such a long sequence prevents effectively training neural models and fitting GPU memory because of sequence length limits. For example, the BERT (Devlin et al., 2019) has a limitation of 512 tokens. We solve the problem by splitting each document longer than a length limit into multiple snippets with a random size, which is between 200 and 512 tokens. This simulates clinical settings in real-world that physicians can recognize their patients with partial symptom descriptions.

3.3. Task 2: Patient-Concept

The second task is to *enrich* user representations by extracted concepts from clinical notes. On the one hand, concepts are identities of users if the concepts co-occur with the users. On the other hand, concept entities can connect users who share similar medical symptoms and separate users who do not share similar information. Given that C_u be a list of extracted medical concepts from the patient's clinical notes and we have $c \in C$, following task 1, we can formalize our goal as maximizing the conditional probability of $P(C_u|u)$. Let $c_{u,k}$ as a medical concept (index as k) that occurs in the clinical notes of the patient u, our patient-document task aims to maximize the following probability:

$$P(C_u|u) = \sum_{k=1}^{|C_u|} log P(c_{u,k}|u) \qquad (3)$$

To reduce computational complexity, we employ negative sampling to approximate our task goal. The negative sampling can help us convert the task to a binary prediction task that checks if a sampled concept exists in the patient's clinical notes. If a sampled concept exists in the target patient's clinical notes, the label is 1; otherwise, the prediction label is 0. The classification problem minimizes loss values by the binary cross-entropy:

$$\begin{aligned} \mathcal{L}(u,c) = &-log(\sigma(v(u) \cdot v(c_u))) \\ &- log(1 - \sigma(v(u) \cdot v(c_{\tilde{u}}))) \end{aligned} \qquad (4)$$

, where $c_{\tilde{u}}$ refers to negative samples of noise documents. Following the task 1, we use the sigmoid function as our non-linear function to normalize values of the dot production. For each concept, we sample another 3 noise concept sets which do not co-occur with the target patient.

3.4. Joint Learning

We optimize user embeddings by patient-document, patient-concept, and masked language modeling (mlm) tasks in parallel. The joint tasks can incorporate document and concept information into user representations. For the mlm, we follow pre-training tasks of the BERT (Devlin et al., 2019) to randomly masked out 15% tokens *only if* the CAUE models are using the BERT as neural feature extractors. We sum the total loss value by:

$$\mathcal{L} = \lambda\mathcal{L}(u,c) + (1-\lambda)\mathcal{L}(u,d) + \alpha\mathcal{L}_{mlm} \qquad (5)$$

where λ and α are weight factors to balance different tasks. We empirically set λ as 0.3 and α as 0.03.

4. Experiments

In this section, we present experimental settings of our approach and baselines. To keep consistency, we fix embedding dimensions as 300 for user and concept embeddings. We include more implementation details in our appendix.

4.1. CAUE Settings

We treat neural models as feature extractors and experiment with two types of neural document encoders, Gated Recurrent Unit (GRU) (Cho et al., 2014) and

| Models | Phenotype Prediction | | Mortality | Patient Relatedness | | Retrieval | |
	Diabetes	MIMIC-III	MIMIC-III	Diabetes	MIMIC-III	Diabetes	MIMIC-III
Baselines							
word2user	0.574	0.321	0.861	0.530	0.809	0.374	0.065
doc2user	0.288	0.235	0.425	0.573	0.253	0.344	0.132
dp2user	0.416	0.315	**0.868**	0.542	0.829	0.318	0.113
suisil2user	0.413	0.274	0.424	0.542	0.830	0.313	0.132
usr2vec	0.556	0.158	0.423	0.355	0.252	0.349	0.137
siamese2user	0.590	0.253	0.750	0.410	0.218	0.360	0.116
Ours							
CAUE-GRU	**0.635**	**0.369**	0.817	**0.292**	**0.205**	0.386	0.139
CAUE-BERT	0.593	0.339	0.792	0.337	0.235	**0.396**	**0.142**

Table 2: Two types of performance evaluations: supervised (phenotype and mortality classification) and unsupervised (patient relatedness and retrieval). For the patient relatedness, lower scores are better.

BERT (Devlin et al., 2019). We denote the two types of encoders as *CAUE-GRU* and *CAUE-BERT* in the following. Each type of encoder reads through sequential tokens and generates document representations. We set maximum epochs as 15, batch size as 16, and vocabulary size as 15,000 for both token and concept.

CAUE-GRU utilizes one GRU layer to encode each document. To incorporate more contextual information, we employ a bi-directional GRU layer. Its output is a concatenation of bi-directional hidden vectors. We set the dimension of the GRU layer as 300 and apply a dropout with 0.2 on outputs of the GRU layer. We use a pre-trained word embedding (Zhang et al., 2019) to encode tokens into vectors. The embedding was pre-trained on PubMed and MIMIC-III corpora. While extensive research methods have explored how to train concept embeddings (Choi et al., 2016; Zhang et al., 2020), our study uses the concepts as mediums to learn hidden patterns to better separate patients. We induce an initial concept embedding for (multi-token) concepts by averaging constituent word vectors. We optimize the model using RMSprop (Hinton et al., 2012) with a learning rate $1e - 4$.

CAUE-BERT encodes document into fixed-length vectors via BERT (Devlin et al., 2019). The BERT model encodes each document by a hierarchical transformer (Vaswani et al., 2017) layers. We add a fully connected layer on top of the encoded `[CLS]` token from the last layer of the BERT model to map 768-dimensional vectors into 300 dimensions. To get document vectors, we pad input documents into a fixed-length, which is a hyperparameter. As initializing the concept embedding with other pre-trained embed-

dings does not align with the BERT in the vector space, we treat concepts as short sentences and utilize the BERT model to obtain concept embeddings. We experimented with multiple pre-trained BERT models (Peng et al., 2019; Alsentzer et al., 2019; Lee et al., 2019). However, we did not find significant improvements with the different choices. We adopt AdaW (Loshchilov and Hutter, 2019) to optimize model parameters with a learning rate of $3e - 5$. We pad documents with less than 512 tokens to ensure every input document has an equal size.

4.2. Baselines

In this study, we compare our approach with six unsupervised user embeddings. To keep fair comparison, we keep the same embedding dimensions, run experiments with three times, and average model performances. We use the same pre-trained embeddings (Zhang et al., 2019; Alsentzer et al., 2019) throughout the baselines.

word2user represents users by averaging over aggregated word representations (Benton et al., 2016). We calculate a user representation by averaging embeddings of all tokens in the clinical notes of each patient. We use the pre-trained word embedding (Zhang et al., 2019) to encode tokens.

doc2user applies paragraph2vec (Le and Mikolov, 2014) to obtain user vectors. We implemented the User-D-DBOW model, which achieved the best performance in the previous work (Ding et al., 2018). The model trains the paragraph2vec for 10 epochs and left other parameters with default values in the

Gensim (Rehurek and Sojka, 2010). We append clinical documents of each patient as a single document. Then the User-D-DBOW model can derive a single user vector from the aggregated document.

suisil2user combines the doc2user with medical concepts to obtain patient vectors (Sushil et al., 2018). We reuse doc2vec to derive document features and follow the previous work to extract concept features. We first derive TF-IDF weighted scores of medical concepts for each patient and train an autoencoder to reduce the dimensions of concept features. The autoencoder learns user embeddings by compressing features inputs into intermediate embeddings and reconstructing them back feature inputs. Following the previous work, we set the sigmoid as the default activation function. We add a 0.05 noise rate and train the autoencoder for ten epochs. The document vectors concatenate with dimension-reduced concept vectors to generate patient representations.

dp2user generates patient representations by applying an autoencoder on extracted topic features (Miotto et al., 2016). We train Latent Dirichlet Allocation (LDA) (Blei et al., 2003) on the patient level and apply the LDA on clinical notes to derive topic features. We set the number of topics as 300, train the model for 10 epochs, and leave the rest of the parameters as their default values in Gensim (Rehurek and Sojka, 2010). We train a four-layer autoencoder with a noise score of 0.05 to reduce dimensions of topic features.

usr2vec trains user embeddings by predicting if users authored sampled tokens (Amir et al., 2017). The predictive goal is to measure if sampled words co-occur with a patient in clinical notes in terms of clinical notes. To make a fair comparison, instead of randomly initializing word embeddings from scratch, we initialized embedding weights as the same pre-trained embeddings (Zhang et al., 2019) in our approach. We train the model for 10 epochs, set the number of negative samples as three.

siamese2user utilizes a Siamese (Mueller and Thyagarajan, 2016) recurrent architecture (Huang et al., 2021) that learns patient representations on the document level. The Siamese model measures similarities between two sentences. We encode each clinical note via a bidirectional GRU, which is beyond the token-level usr2vec. For the patient, we apply a feed-forward network on the patient vectors and optimize the model by distances between document and patient vectors.

4.3. Extrinsic 1: Phenotype prediction

We train classifiers with user embeddings as input features to infer phenotype (ICD-9 codes). Each patient has an encoded set, a one-hot vector, which is a length of the unique labels. The task is to evaluate how accurately a classifier can predict the clinical labels of each patient. We use the embedding to encode patients and feed patient vectors to the logistic regression (LR) classifier for each user embedding. We optimize the LR classifier by Adam (Kingma and Ba, 2014) and L2-regularization. We set the learning rate as 0.001 and the regularization score as 0.01. We use 5-fold cross-validation to evaluate classification performance. Each fold splits the Diabetes and MIMIC-III data based on their patient counts by 80% as a training set and 20% as a test set. Finally, we report mean average precision (MAP) for model evaluation.

We present our classification results in Table 2. Comparing to the baselines, LR classifiers trained by our proposed approach (CAUE-GRU) achieve the best performance across the two datasets. The results highlight that our method can capture patient variations and augment phenotype inference via the joint optimization tasks. The only difference between the siamese2user and usr2vec methods is whether it is on document level. We can find that siamese2user outperforms usr2vec, which indicates that learning unsupervised user embedding can benefit from sequential dependency beyond individual tokens.

4.4. Extrinsic 2: In-hospital mortality prediction

In-hospital mortality is a binary classification task to predict if patients died during in-hospital stay based on the first 48 hours of an ICU stay. We follow the benchmark (Harutyunyan et al., 2019) to generate inpatient mortality labels. The mortality annotations are only available for the MIMIC-III dataset. We use 5-folds cross-validation to train LR classifiers and F1-score (macro) to evaluate the performance of mortality prediction as for the macro is less sensitive to imbalanced mortality labels.

We present mortality prediction results in Table 2. Our approach generally performs well among baselines. We can find that our close work usr2vec performs worse than siamese2user. Training mechanisms of usr2vec is the poor performance on both classification tasks. The usr2vec predicts relations between words and users on Twitter messages, which have much shorter documents. The negative sampling space will

be too sparse to find strong patterns that separate patients. This highlights the unique challenge to derive unsupervised user embeddings on clinical settings.

4.5. Intrinsic 3: patient relatedness

The patient relatedness task is to measure semantic relation or similarity between patients. Data to evaluate relatedness originates from lexical semantics. SimLex-999 (Hill et al., 2015) provides a list of word pairs with similarity scores from human annotators. SimLex-999 checks if similarities of word vectors are close to human perceptions. Inspired by this, we utilize an intrinsic evaluation that measures the quality of user embeddings by patient relatedness. However, unlike word embedding evaluation, there is no human judgement for patient-level similarities. Therefore, we adapt medical codes to measure patient similarities. The patient relatedness assumes that the similarity between every two patients should be proportional to the similarity of their clinical labels:

$$Similarity(u_1, u_2) \propto Similarity(l_1, l_2) \quad (6)$$

, where $u1 \in U$, $u2 \in U$, l_1 and l_2 are patients' corresponding phenotype labels. We use user embeddings to encode patients into vectors and encode the labels into one-hot vectors. The dimension of each label vector is the number of unique clinical labels. For example, l_1 in MIMIC-III data has 276 dimensions, corresponding to number of unique phenotype labels in Table 1. We use cosine similarity to measure patient relatedness between vectors. *Mean Squared Error* (MSE) is to measure differences between a pair of patients and a pair of label vectors:

$$MSE = \frac{1}{|U|} \sum_{i=1}^{|U|} (Sim(l_1, l_2) - Sim(u_1, u_2))^2 \quad (7)$$

We present our evaluation results of patient relatedness in Table 2. Overall our concept-aware user embedding (CAUE-GRU) achieves the best performance over baselines. While the suisil2user achieves better performance than the doc2user on Diabetes, it fails on the MIMIC-III. Furthermore, our approach yields good performance on the two datasets. The result can highlight that how to encode concept features into user embeddings are critical for improving model robustness.

4.6. Intrinsic 4: patient retrieval

The retrieval task is to treat each patient as a "query" and measure relations between the query and its retrieved top ranks. We find the top 10 most similar patients for a query patient by calculating cosine similarities between the query patient vector and other patient vectors. To measure the effectiveness of user embeddings, we need to examine the quality of top-ranking returns. Intuitively, the top returns should share similar clinical conditions with the query patient. However, clinical labels of each patient can be more than the binary labels in previous work (Amir et al., 2017) or in-hospital mortality. The query aims to find the best candidates so that similarities between the query and the other candidates should be close to 1. Therefore, we adapt *jaccard coefficient* to measure similarity between a query patient (u_1) and a retrieved patient (u_2) by their clinical labels:

$$Jaccard(u_1, u_2) = \frac{|l_1 \cap l_2|}{|l_1 \cup l_2|} \quad (8)$$

We choose ten top ranks and evaluate retrieved candidates by the average Jaccard coefficients. We present evaluation results of the retrieval task in Table 2. The CAUE-GRU outperforms baselines by a large margin. The result highlights that modeling language sequential information can be beneficial to learn patient clinical conditions for retrieval tasks. The CAUE outperforms the siamese2user, which does not incorporate concepts. Performance improvements show that incorporating medical concepts and leveraging document-level relations can significantly boost model performance.

4.7. Effectiveness of Medical Concepts

We now investigate the effectiveness of medical concepts in more detail. While our approach uses neural concept embeddings, other baselines do not. Our main goal is to examine if medical concepts can generally improve the performance of user representations.

We update five baselines (word2user, doc2user, dp2user, usr2vec, and siamese2user) to incorporate concept information into training user embedding under the multitask learning framework. For the *word2user*, the model aggregates all tokens of concepts, average over the concept vectors, and combine both token and concept vectors to obtain new patient representations. For the *doc2user*, we treat groups of medical concepts as documents, build a paragraph2vec (Le and Mikolov, 2014) model to convert

+ Concept	Phenotype Prediction		Mortality	Patient Relatedness		Retrieval	
	Diabetes	MIMIC-III	MIMIC-III	Diabetes	MIMIC-III	Diabetes	MIMIC-III
word2user	+7.6% (.044)	+40.4% (.130)	+4.8% (.041)	+3.4% (-.018)	+1.2% (-.010)	-0.5% (-.002)	+24.6% (.016)
doc2user	-43.4% (-.125)	-32.3% (-.076)	+2.8% (.012)	+1.1% (-.006)	+0.8% (-.02)	+6.1% (.021)	-3.0% (-.004)
dp2user	+22.6% (.094)	+52.1% (.164)	+3.8% (.033)	+2.0% (-.011)	+24.9% (-.207)	+17.6% (.056)	+18.6% (.021)
usr2vec	+9.4% (.052)	+249.0% (.394)	+111.1% (.470)	+6.2% (-.022)	-63.7% (.444)	+9.5% (.033)	-36.5% (-.050)
siamese2user	+7.6% (.045)	+45.8% (.116)	+45.8% (.116)	+28.7% (-.118)	+6.0% (-.013)	+7.2% (.026)	+19.8% (.023)
Average	+.8% (.013)	+61.84 (.146)	+33.66% (.134)	+8.28% (-.035)	-6.16% (.194)	+8.0% (.027)	+4.7 (.001)
Median	+9.4% (.052)	+45.8% (.116)	+4.8% (.041)	+3.4% (-.018)	+1.2% (-.010)	+7.2% (.026)	+18.6% (.021)

Table 3: Performance gains of user embedding models combining with medical concepts (+Concept) comparing to standard non-concept information. We report percentage increases with absolute value increases in brackets after incorporating medical concepts. For patient relatedness, because lower is better, a negative value within brackets means a performance gain.

concepts into vectors, and concatenate both tokens with the new concept vectors. For the *dp2user*, we build a second topic model on medical concepts besides the first topic model over tokens. We then derive concept vectors by the second topic model and train an autoencoder to reduce dimensions of the concept-topic features. Finally, we combine the compressed concept and token vectors from the two separated autoencoders. For the *siamese2user*, we build a second prediction task to compare distances between user and concept embeddings. For the *usr2vec*, we create two joint predictions, user-token and user-concept predictions. Both prediction tasks follow the negative sampling process to learn user embeddings. Next, we initialize concept embeddings by averaging vectors of tokens within each concept, which gets token vectors from the pre-trained word embedding (Zhang et al., 2019).

Table 3 shows the percentage improvement and absolute value increases in brackets across four evaluation tasks when incorporating medical concepts compared to baselines without explicitly leveraging medical concepts. Overall, medical concepts generally augment the performance of user embeddings with an average absolute improvement of 5.93% on Diabetes and 23.51% on MIMIC-III respectively. The medical concepts appear to be particularly important for CAUE, improving performance on all four evaluation tasks with an average increase in performance up to 20.3%. Our close work usr2vec gains a significant improvement with an average of 40.7%. Comparing the different methods for incorporating medical concepts, on the one hand, our proposed CAUE-GRU works the best on average overall; on the other hand, joint learning of concepts and clinical notes can generally improve effectiveness of the baselines.

5. Related Work

In this work, we focus on learning user embeddings from **clinical notes**, while some studies (Lei et al., 2018; Wang et al., 2019; Yin et al., 2019; Steinberg et al., 2021) learn patient representations from tabular data. Yin et al. applies tensor factorization to learn patient representations from tabular data, including diagnosis codes and laboratory tests. However, learning user embeddings from clinical notes is our focus in this study, and clinical notes are noisy and unstructured data. Several studies have proposed supervised user embeddings for clinical notes (Wu et al., 2020; Si and Roberts, 2019). Dligach and Miller learn patient vectors by feed-forward neural networks on concept embeddings. The model averages concept vectors for each patient and predicts ICD-9 codes (Dligach and Miller, 2018). Si and Roberts presents a hierarchical recurrent neural network to train embeddings for patients. The method trains a model on MIMIC-III with phenotype annotations and applies the pre-trained model on other tasks (Si and Roberts, 2020). However, unsupervised user embeddings do not have medical annotations to learn hidden structures of patients. Autoencoder is an encoder-decode neural network that takes patient features as inputs, compresses the features into fixed-length vectors, and reconstructs the vectors back to its input features (Miotto et al., 2016; Sushil et al., 2018). However, the autoencoder-based approaches require well-designed feature sets, and the dimension ratio between input features and user embeddings may reduce training effectiveness. Our study fills the gap by proposing an unsupervised user embedding under a multitask framework that enforces models to recognize patients of clinical notes. The approach does not require any human supervision and can learn user embeddings in an end-to-end style.

Concept embeddings aim to learn vector representations for medical concepts. Research treats medical concepts as tokens or phrases and encodes them into vectors by contextualized word embeddings (Zhang et al., 2020), medical ontology (Song et al., 2019) and graphical neural networks (Choi et al., 2016, 2020). However, few studies have incorporate medical concepts into learning patient representations under unsupervised settings. Our study jointly leverages clinical notes and medical concepts to learn user embeddings for patients. While our study does not focus on concept embeddings, we utilize the concept embeddings to enrich patient representations. Recent studies utilized concepts in multimodal machine learning to align embedding spaces between vision and language (Zhang et al., 2021). Our study utilizes concepts as additional constraints that balance semantic relations between patient, clinical text descriptions, and medical concepts.

Multitask learning simultaneously trains a model with multiple tasks. For example, Si and Roberts optimize patient representations by two supervised tasks, mortality and length of stay predictions (Si and Roberts, 2019). However, this study handles a triangle relation across patient, clinical notes, and medical concepts. We define the relation predictions as tasks to joint optimize patient representations instead of the downstream tasks. Our approach trains user embeddings from multiple sources and uses medical concepts to enrich patient representations. While studies (Miotto et al., 2016; Sushil et al., 2018) utilize features from different sources, they concatenate patient features and train models in a single task.

6. Conclusion

In this study, we have proposed a concept-aware user embedding (CAUE) in an unsupervised setting. The model jointly leverages information of medical concepts and clinical notes using multitask learning. The CAUE is an end-to-end neural model without any human supervision via contrastive learning and negative sampling. Our extrinsic and intrinsic evaluations demonstrate the effectiveness of our approach, which outperforms baselines by a large margin. Nonetheless, the medical concept shows its effectiveness in improving unsupervised baselines in general. We include additional implementation details in the appendix.

Limitation. While experiments have demonstrated the effectiveness of our proposed unsupervised approach in clinical settings, several limitations must be acknowledged to interpret our findings appropriately. *First*, user behaviors can shift over time (Han et al., 2020). In this study, we have proposed methods to isolate data entries by a patient per hospital visit. The strategy can effectively reduce the temporal impacts on modeling and evaluation. *Second*, data sources may hurdle model evaluations due to their different types of clinical annotations. We keep evaluations consistent across the two datasets and conduct intrinsic evaluations to avoid training a large number of classifier parameters.

Institutional Review Board (IRB)

This study has no human-subject research and only uses publicly available and de-identified data, which does not need an IRB approval.

Acknowledgments

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1909. URL https://www.aclweb.org/anthology/W19-1909.

Silvio Amir, Glen Coppersmith, Paula Carvalho, Mario J. Silva, and Bryon C. Wallace. Quantifying mental health from social media with neural user embeddings. In Finale Doshi-Velez, Jim Fackler, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 306–321, Boston, Massachusetts, 18–19 Aug 2017. PMLR. URL http://proceedings.mlr.press/v68/amir17a.html.

Alan R Aronson and François-Michel Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 05 2010. ISSN 1067-5027. doi: 10.1136/jamia.2009.002733. URL https://doi.org/10.1136/jamia.2009.002733.

Adrian Benton, Raman Arora, and Mark Dredze. Learning multiview embeddings of twitter users. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-2003. URL https://www.aclweb.org/anthology/P16-2003.

Steven Bird and Edward Loper. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, July 2004. URL https://www.aclweb.org/anthology/P04-3031.

Elena Birman-Deych, Amy D Waterman, Yan Yan, David S Nilasena, Martha J Radford, and Brian F Gage. Accuracy of icd-9-cm codes for identifying cardiovascular and stroke risk factors. *Medical care*, 43 (5):480–485, 2005. URL http://www.jstor.org/stable/3768402.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. URL http://www.jmlr.org/papers/v3/blei03a.html.

Olivier Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(Database issue):D267–D270, jan 2004. ISSN 1362-4962. doi: 10.1093/nar/gkh061. URL https://pubmed.ncbi.nlm.nih.gov/14681409https://www.ncbi.nlm.nih.gov/pmc/articles/PMC308795/.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22243–22255. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/fcbc95ccdd551da181207c0c1400c655-Paper.pdf.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-4012. URL https://www.aclweb.org/anthology/W14-4012.

Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. Multi-layer representation learning for medical concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1495–1504, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939823. URL https://doi.org/10.1145/2939672.2939823.

Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. Gram: Graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 787–795, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348874. doi: 10.1145/3097983.3098126. URL https://doi.org/10.1145/3097983.3098126.

Edward Choi, Zhen Xu, Yujia Li, Michael Dusenberry, Gerardo Flores, Emily Xue, and Andrew Dai. Learning the graphical structure of electronic health records with graph convolutional transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 606–613, New York, US, Apr. 2020. AAAI. doi: 10.1609/aaai.v34i01.5400. URL https://ojs.aaai.org/index.php/AAAI/article/view/5400.

S. Darabi, M. Kachuee, S. Fazeli, and M. Sarrafzadeh. Taper: Time-aware patient ehr representation. *IEEE Journal of Biomedical and Health Informatics*, 24(11):3268–3275, 2020. doi: 10.1109/JBHI.2020.2984931. URL https://doi.org/10.1109/jbhi.2020.2984931.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, June 2019. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

T. Ding, W. K. Bickel, and S. Pan. Predicting delay discounting from social media likes with unsupervised feature learning. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 254–257, Barcelona, Spain, Aug 2018. IEEE. doi: 10.1109/ASONAM.2018.8508277.

Dmitriy Dligach and Timothy Miller. Learning Patient Representations from Text. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 119–123, New Orleans, Louisiana, jun 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-2014. URL https://www.aclweb.org/anthology/S18-2014.

Lei Han, Alessandro Checco, Djellel Difallah, Gianluca Demartini, and Shazia Sadiq. Modelling user behavior dynamics with embeddings. In *Proceedings of the 29th ACM International Conference on Information; Knowledge Management*, CIKM '20, page 445–454, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368599. doi: 10.1145/3340531.3411985. URL https://doi.org/10.1145/3340531.3411985.

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96, 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0103-9. URL https://doi.org/10.1038/s41597-019-0103-9.

Felix Hill, Roi Reichart, and Anna Korhonen. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, December 2015. doi: 10.1162/COLI_a_00237. URL https://www.aclweb.org/anthology/J15-4004.

Geoffrey E. Hinton, Nitish Srivastava, and Kevin Swersky. Lecture 6a- overview of mini-batch gradient descent, 2012. URL http://www.cs.toronto.edu/{~}tijmen/csc321/slides/lecture{_}slides{_}lec6.pdf.

Xiaolei Huang, Michael J. Paul, Franck Dernoncourt, Robin Burke, and Mark Dredze. User factor adaptation for user embedding via multitask learning. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 172–182, Kyiv, Ukraine, April 2021. URL https://www.aclweb.org/anthology/2021.adaptnlp-1.18.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Liwei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016. doi: 10.1038/sdata.2016.35. URL https://doi.org/10.1038/sdata.2016.35.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, pages 1–15, San Diego, US, 2014. ICLR. URL http://arxiv.org/abs/1412.6980.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In Eric P. Xing and Tony Jebara, editors, *Proceedings of Machine Learning Research*, volume 32, pages 1188–1196, Bejing, China, 22–24 Jun 2014. PMLR. URL http://proceedings.mlr.press/v32/le14.html.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 09 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz682. URL https://doi.org/10.1093/bioinformatics/btz682.

Liqi Lei, Yangming Zhou, Jie Zhai, Le Zhang, Zhijia Fang, Ping He, and Ju Gao. An effective patient representation learning for time-series prediction tasks based on ehrs. In *2018 IEEE International Conference on Bioinformatics and Biomedicine*, pages 885–892, Madrid, Spain, 2018. IEEE, IEEE. doi: 10.1109/BIBM.2018.8621542. URL https://doi.org/10.1109/BIBM.2018.8621542.

Sheng Li and Handong Zhao. A survey on representation learning for user modeling. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4997–5003, online, 7 2020. International Joint Conferences on Artificial Intelligence Organization. doi: 10.24963/

ijcai.2020/695. URL https://doi.org/10.24963/ijcai.2020/695. Survey track.

Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJvJXZb0W.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA, 2013. Curran Associates Inc. URL http://dl.acm.org/citation.cfm?id=2999792.2999959.

Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6(1):1–10, 2016. doi: 10.1038/srep26094. URL https://doi.org/10.1038/srep26094.

Jonas Mueller and Aditya Thyagarajan. Siamese recurrent architectures for learning sentence similarity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, Mar. 2016. URL https://ojs.aaai.org/index.php/AAAI/article/view/10350.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, June 2018. doi: 10.18653/v1/N18-1100. URL https://www.aclweb.org/anthology/N18-1100.

Shimei Pan and Tao Ding. Social media-based user embedding: A literature review. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6318–6324, Macao, China, 7 2019. International Joint Conferences on Artificial Intelligence Organization.

doi: 10.24963/ijcai.2019/881. URL https://doi.org/10.24963/ijcai.2019/881.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL https://arxiv.org/pdf/1912.01703.pdf.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011. ISSN 15324435.

Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, August 2019. doi: 10.18653/v1/W19-5006. URL https://www.aclweb.org/anthology/W19-5006.

Radim Rehurek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, may 2010. ELRA. ISBN 2-9517408-6-7. URL https://radimrehurek.com/gensim/lrec2010_final.pdf.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 09 2010. ISSN 1067-5027. doi: 10.1136/jamia.2009.001560. URL https://doi.org/10.1136/jamia.2009.001560.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307, September 2015. doi: 10.18653/v1/D15-1036. URL https://www.aclweb.org/anthology/D15-1036.

Yuqi Si and Kirk Roberts. Deep patient representation of clinical notes via multi-task learning for mortality prediction. *AMIA Summits on Translational Science Proceedings*, 2019:779, 2019. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6568068/.

Yuqi Si and Kirk Roberts. Patient representation transfer learning from clinical notes based on hierarchical attention network. *AMIA Summits on Translational Science Proceedings*, 2020:597, 2020. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7233035/.

Yuqi Si, Jingcheng Du, Zhao Li, Xiaoqian Jiang, Timothy Miller, Fei Wang, W. Jim Zheng, and Kirk Roberts. Deep representation learning of patient data from electronic health records (ehr): A systematic review. *Journal of Biomedical Informatics*, 115:103671, 2020. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2020.103671. URL https://www.sciencedirect.com/science/article/pii/S1532046420302999.

Lihong Song, Chin Wang Cheong, Kejing Yin, William K. Cheung, Benjamin C. M. Fung, and Jonathan Poon. Medical concept embedding with multiple ontological representations. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4613–4619. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/641. URL https://doi.org/10.24963/ijcai.2019/641.

Ergin Soysal, Jingqi Wang, Min Jiang, Yonghui Wu, Serguei Pakhomov, Hongfang Liu, and Hua Xu. CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association*, 25(3):331–336, 11 2017. ISSN 1527-974X. doi: 10.1093/jamia/ocx132. URL https://doi.org/10.1093/jamia/ocx132.

Ethan Steinberg, Ken Jung, Jason A Fries, Conor K Corbin, Stephen R Pfohl, and Nigam H Shah. Language models are an effective representation learning technique for electronic health record data. *Journal of Biomedical Informatics*, 113:103637, 2021. ISSN 1532-0464. doi: 10.1016/j.jbi.2020.103637. URL https://www.sciencedirect.com/science/article/pii/S1532046420302653.

Amber Stubbs, Michele Filannino, Ergin Soysal, Samuel Henry, and Özlem Uzuner. Cohort selection for clinical trials: n2c2 2018 shared task track 1. *Journal of the American Medical Informatics Association*, 26(11):1163–1171, 09 2019. ISSN 1527-974X. doi: 10.1093/jamia/ocz163. URL https://doi.org/10.1093/jamia/ocz163.

Madhumita Sushil, Simon Šuster, Kim Luyckx, and Walter Daelemans. Patient representation learning and interpretable evaluation using clinical notes. *Journal of Biomedical Informatics*, 84:103–113, 2018. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2018.06.016. URL https://www.sciencedirect.com/science/article/pii/S1532046418301266.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, Long Beach, CA, US, 2017. NIPS.

Ying Wang, Xiao Xu, Tao Jin, Xiang Li, Guotong Xie, and Jianmin Wang. Inpatient2vec: Medical representation learning for inpatients. In *2019 IEEE International Conference on Bioinformatics and Biomedicine*, pages 1113–1117, San Diego, CA, USA, 2019. IEEE. doi: 10.1109/BIBM47256.2019.8983281. URL https://doi.org/10.1109/BIBM47256.2019.8983281.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, October 2020. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Stephen Wu, Kirk Roberts, Surabhi Datta, Jingcheng Du, Zongcheng Ji, Yuqi Si, Sarvesh Soni, Qiong Wang, Qiang Wei, Yang Xiang, et al. Deep learning in clinical natural language processing: a methodical review. *Journal of the American Medical Informatics Association*, 27(3):457–470, 2020. doi: 10.1093/jamia/ocz200. URL https://doi.org/10.1093/jamia/ocz200.

Kejing Yin, Dong Qian, William K. Cheung, Benjamin C. M. Fung, and Jonathan Poon. Learning phenotypes and dynamic patient representations via rnn regularized collective non-negative tensor factorization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1246–1253, Jul. 2019. doi: 10.1609/aaai.v33i01.33011246. URL https://ojs.aaai.org/index.php/AAAI/article/view/3920.

Bowen Zhang, Hexiang Hu, Linlu Qiu, Peter Shaw, and Fei Sha. Visually grounded concept composition. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 201–215, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.findings-emnlp.20.

Xiao Zhang, Dejing Dou, and Ji Wu. Learning conceptual-contextual embeddings for medical text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9579–9586, New York, US, Apr. 2020. AAAI. doi: 10.1609/aaai.v34i05.6504. URL https://ojs.aaai.org/index.php/AAAI/article/view/6504.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific data*, 6(1):1–9, 2019. doi: 10.1038/s41597-019-0055-0. URL https://doi.org/10.1038/s41597-019-0055-0.

Appendix A. Ablation Analysis

We compare performance impacts of individual modules in our proposed approach. The comparison experiments follow the previous experimental settings in Section 4 and trained models with different learning rates within $[1e-4, 1e-6]$. We experiment with the two base neural encoders, GRU and BERT, in our CAUE framework with both the contrastive learning (CT) and medical concepts (CC). There are four combination types on top of the base neural feature extractors (GRU and BERT): 1) none of the components (None), 2) only with CT (+CT), 3) only with CC (+CC), 4) and with both optimization components (+CC+CT). We evaluate the four combinations on the four tasks, phenotype prediction, mortality, patient relatedness, and retrieval.

We present results in Table 4. Generally, learning user embeddings with CT and CC can achieve the best performance across the evaluation tasks. Both CT and CC can improve the embedding performance. This indicates that the two proposed strategies work coherently to improve model effectiveness on extracting patient information from clinical notes and medical concepts. And we can also observe that medical concepts generally have more improvements over the contrastive learning. For example, +CC models usually outperforms +CT models. This indicates that medical concepts may have more indicative information and correlations with the clinical codes.

Appendix B. Implementation Details

We follow previous work (Mullenbach et al., 2018; Harutyunyan et al., 2019; Stubbs et al., 2019) to process MIMIC-III data. Harutyunyan et al. provides methods to calculate patient age information, and we wrote regular expressions to extract age information from the Diabetes corpus. We keep clinical records of adults (≥ 18) in both MIMIC-III and Diabetes datasets. The Diabetes data aggregates discharge summaries of each patient into one file, and we split individual discharge summaries by its record separator. To keep consistent, we use discharge summaries of both data. We implement our models and re-implement baselines using PyTorch (Paszke et al., 2019), Huggingface Transformers (Wolf et al., 2020), scikit-learn (Pedregosa et al., 2011) and NLTK (Bird and Loper, 2004). For the CAUE-BERT, we experiment with multiple pre-trained BERT models (Peng et al., 2019; Alsentzer et al., 2019; Lee et al., 2019). However, we did not find significant improvements with the different choices. Initially, we trained models of Diabetes and MIMIC-III datasets on a NVIDIA GeForce RTX 3090, and conducted evaluation experiments on CPUs. For the baselines, we train user2vec on the GPU and other baselines on CPUs. We also experimented different learning rates within $[1e-4, 1e-6]$ and trained models on the university HPC cluster with NVIDIA Tesla V100s. We then conducted the ablation analysis and reported results of the experiments.

CAUE		Phenotype Prediction		Mortality	Patient Relatedness		Retrieval	
		Diabetes	MIMIC-III	MIMIC-III	Diabetes	MIMIC-III	Diabetes	MIMIC-III
GRU	None	0.573	0.191	0.658	0.364	0.231	0.359	0.119
	+CT	0.587	0.284	0.691	0.334	0.226	0.363	0.106
	+CC	0.618	0.348	0.814	0.319	0.207	0.377	0.131
	+CC+CT	0.635	0.369	0.817	0.292	0.205	0.386	0.139
BERT	None	0.545	0.197	0.648	0.517	0.262	0.322	0.101
	+CT	0.569	0.198	0.660	0.503	0.244	0.351	0.103
	+CC	0.586	0.242	0.692	0.419	0.250	0.382	0.126
	+CC+CT	0.593	0.339	0.792	0.337	0.235	0.396	0.142

Table 4: Ablation study performance. None indicates no contrastive learning (CT) and medical concept (CC). The + means adding the training component. While lower scores for the patient relatedness are better, higher scores for the other tasks are better.

Appendix C. Exploratory Analysis of Medical Concepts

Phenotype inference and patient retrieval are essential tasks for diagnosis and knowledge discovery. Learning patient representations is critical for building robust models. Medical concepts composed of one or more tokens in various combinations can indicate diverse semantic meanings than individual tokens. For example, separations of "bid" and "protein" can have different semantic meanings than a combination of them, which is one type of proteins during cell death receptormediated apoptosis. Medical concepts can correlate closely with medical annotations. For example, "drinker" indicates people who drink alcohol, while "heavy drinker" identified in Diabetes data frequently co-occurs with the cohort annotation, "alcohol-abuse". This section aims to test how the extracted concepts in our particular datasets can contribute to user modeling and how strong the effects are.

Dataset	Feature	Coefficient
Diabetes	n-gram	-0.037
	concept	0.388*
MIMIC-III	n-gram	0.405*
	concept	0.146*

Table 5: Multivariate linear regression analysis between feature and label similarities. The variables are two features n-gram and concept. * indicates that p-vale < 0.05.

To answer the question, we conduct a patient retrieval task by treating each patient as a query and retrieve similar patients. We define the similarity between every two patients by calculating a cosine score between their phenotype labels. We assume that if two patients are similar, their phenotype annotations should be similar. Therefore, if document features can accurately describe patients, the feature similarity of two patients should be proportional to their annotation similarity.

We examine the effects of concept features by a multivariate linear regression to examine relations between patient annotation and the two features types. To represent patient features, we derive and combine two feature types, n-gram and concept. We extract uni-, bi-, and tri-gram features and concept features that are normalized by TF-IDF. We calculate feature and medical annotation (ICD-9 codes) similarities between every two patients. Then, we can compare the two feature types by the linear regression between feature and medical annotation similarities.

We show the regression analysis results in Table 5. In both datasets (Diabetes and MIMIC-III), concept features show a significant correlation with medical annotations, while the n-gram features fail the significance test on the Diabetes data. The failed test verifies the null hypothesis that there is no linear correlation between n-gram feature and medical annotation. The results also indicate that incorporating medical concepts could help retrieve similar patients. Concept features show a positive correlation with the medical annotation similarity. The positive correlation means that as the annotation similarity of two patients increases, the concept feature similarity also intends to increase. Those observations suggest there is a strong linear connection between medical concept and phenotype.