Two heads are better than one: Enhancing medical representations by pre-training over structured and unstructured electronic health records

Sicen Liu[1,2], M.S., liusicen@stu.hit.edu.cn

Xiaolong Wang[1], Ph.D., wangxl@insun.hit.edu.cn

Yongshuai Hou[2], Ph.D., houysh@pcl.ac.cn

Ge Li[2,3], Ph.D., geli@pku.edu.cn

Hui Wang[4], M.S., wanghui@gennlife.com

Hui Xu[4], Ph.D., xuhui@gennlife.com

Yang Xiang[2,*], Ph.D., xiangy@pcl.ac.cn

Buzhou Tang[1,2,*], Ph.D., tangbuzhou@gmail.com

[1] Harbin Institute of Technology (Shenzhen), Shenzhen, China

[2] Peng Cheng Laboratory, Shenzhen, China

[3]Peking University, Beijing, China

[4]Gennlife(Beijing) Technology Co Ltd, Beijing, China

*Corresponding authors

Abstract

The massive context of electronic health records (EHRs) has created enormous potentials for improving healthcare, among which structured (coded) data and unstructured (text) data are two important textual modalities. They do not exist in isolation and can complement each other in most real-life clinical scenarios. Most existing researches in medical informatics, however, either only focus on a particular modality or straightforwardly concatenate the information from different modalities, which ignore the interaction and information sharing between them. To address these issues, we proposed a unified deep learning-based medical pre-trained language model, named UMM-PLM, to automatically learn representative features from multimodal EHRs that consist of both structured data and unstructured data. Specifically, we first developed parallel unimodal information representation modules to capture the unimodal-specific characteristic, where unimodal representations were learned from each data source separately. A cross-modal module was further introduced to model the interactions between different modalities. We pre-trained the model on a large EHRs dataset containing both structured data and unstructured data and verified the effectiveness of the model on three downstream clinical tasks, i.e., medication recommendation, 30-day readmission and ICD coding through extensive experiments. The results demonstrate the power of UMM-PLM compared with benchmark methods and state-of-the-art baselines. Analyses show that UMM-PLM can effectively concern with multimodal textual information and has the potential to provide more comprehensive interpretations for clinical decision making.

Introduction

The growing availability of large-scale electronic health records (EHRs) has increased the opportunities of improving healthcare using deep learning methods[1–5]. EHRs usually consist of heterogeneous data records, such as vital signs, medications, laboratory measurements, observations, as well as clinical notes recorded by care practitioners, which carry patients' health status and stages of medical care[6–8]. EHRs often play an important and determinant role for data-driven clinical decision support systems[9–11]. There have been an extensive array of successes achieved in the past few years in deep learning-aided healthcare[12–21], triggering more in-depth analyses on the nature of EHRs to develop effective models[22–26]. Well-organized structured data, together with unstructured data from EHRs, provided a basis for physicians' decision-making[27–30]. These data not only can express informative messages for patients individually but can also be comprehensively overlapped and co-affected[31–33]. A common law to build effective EHR learners, thereby, is to align texts from different modalities to a shared space so as to maximize the utility of the interactive information[34,35]. This also necessitates the deep learning models to be capable of learning unified representations from these multimodal textual EHRs, and understand the modality-specific semantics as well as cross-modal interactions, which might be sustainable for downstream tasks. Figure 1 shows an example of the complementary relationship in a paired piece of multimodal EHR, where the same color indicates a possible overlap between the structured data and the unstructured data.
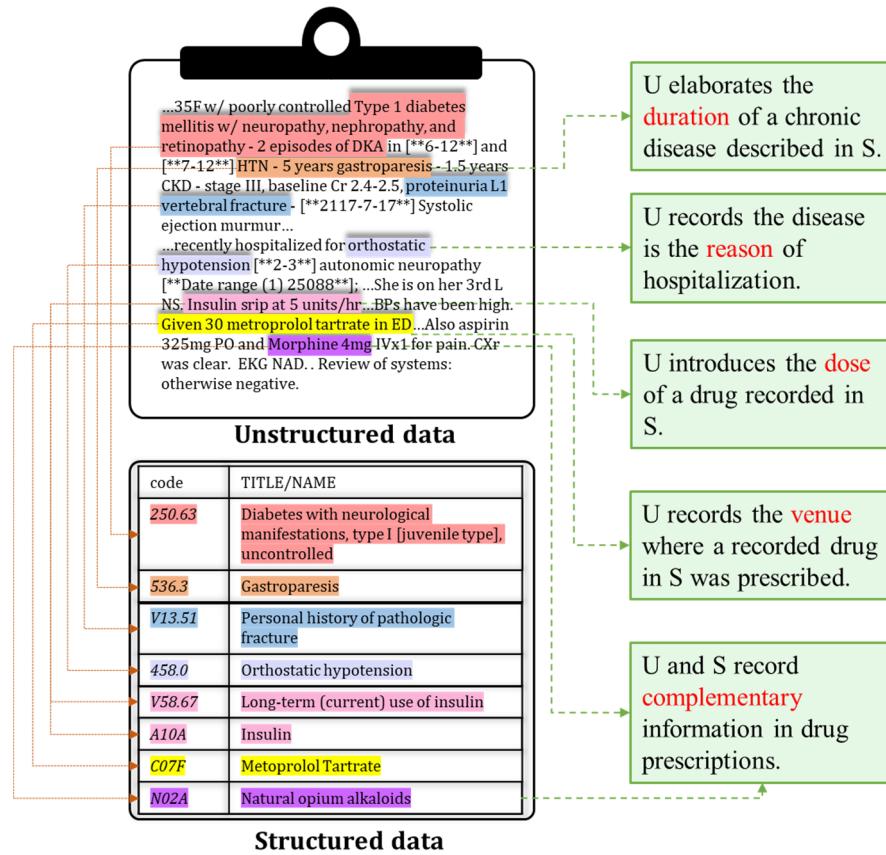
Figure 1. An example of unstructured data and structured data in a visit record of a patient. The structured data show precise and summarized diagnoses and drug names, while the unstructured data records more detailed information such as the duration of a certain disease, the primary reason for hospitalization, and the dosage of a prescription, and shows the complement of the unstructured and structured data. U: unstructured data, S: structured data.

There exist abundant researches for unimodal EHR applications, especially for unstructured data[36–39], which have obtained promising results with the help of pre-trained language models (PLMs)[40–44]. The success of Bidirectional Encoder Representations from Transformers (BERT)[45] in the natural language processing domain has also promoted the transfer of pre-trained models into the medical domain such as the widely used clinicalBERT[46] and BioBERT[42]. For structured data, there are also some preliminary attempts, such as G-BERT[47], BEHRT[48], and Med-BERT[49]. As increasing scenes in clinical applications are not unimodal independent, neither of the aforementioned research paradigms is sufficient in modeling multimodal alignments and interactions. On the other hand, although multimodal learning has been

dramatically developed in recent years, the focus is mainly between images, audios, and texts[31,50–55]. There is an urgent need in developing multimodal methods to jointly model structured data and unstructured data of EHRs.

In this paper, we proposed a **U**nified **M**edical **M**ultimodal **P**re-trained **L**anguage **M**odel (UMM-PLM) for large EHRs, to build the connections between the modalities of structured code and unstructured text. UMM-PLM transforms the multimodal EHRs data into unified representations and uncouples the data from domain-specific tasks. We designed multiple tasks to pre-train UMM-PLM so that downstream clinical tasks could benefit from the unified representation better as prior knowledge. In detail, we firstly developed two unimodal modules for structured data and unstructured data, which can inherit the advantages of unimodal PLMs that were well pre-trained to sustain their modal-specific property. The unimodal modules are simple and flexible that any structured code- and unstructured text-based pre-trained models can be plugged in. We further developed a cross-modal module to model the inter-modality relationships which is composed of a cross-modal attention layer and a modal augment operator. Cross-modal attention is used to learn the complementary information between unstructured data and structured data and model the cross-modal interactions, while the modal augment operator is used to integrating the unimodal-specific and cross-modal features. Two pre-training tasks, i.e. *Text-to-Code* prediction and *Code-to-Code* prediction, are used as the learning objectives of the cross-modal prediction operator.

We evaluated the UMM-PLM model on three medical downstream tasks: medication recommendation, 30-day readmission, and International Classification of Diseases (ICD) coding, which have covered different types of multimodal tasks over EHRs. Different benchmark and state-of-the-art methods were compared through

extensive experiments, and the result demonstrated the effectiveness of the proposed UMM-PLM. Few-shot learning scenarios further show the scalability of the pre-trained model.

Our primary contributions are summarized as follows:

(1) We proposed a novel multimodal pre-trained language model for modeling unstructured data and structured data in EHRs, which can learn cross-modal interactions while retaining unimodal representation abilities;

(2) We conducted fine-tuning experiments on three medical prediction tasks to evaluate the performance of our proposed framework. The results demonstrate the power of the UMM-PLM model in multimodal EHRs;

(3) We conducted ablation studies and results show the effectiveness of each proposed module in UMM-PLM;

(4) Experiments on few-shot learning scenarios demonstrate that our model performs consistently better than baselines;

(5) We have made our codes and pre-trained models publicly available to enhance the reproducibility and offer support to more researchers.

Methods

Data Preparation

We used a large and publicly available database, the Medical Information Mart for Intensive Care III (MIMIC-III) dataset, for our experiments. The MIMIC-III dataset contains data associated with 53,423 distinct hospital admissions for 35,164 adult patients (age 16 years or above) between 2001 and 2012; data includes vital signs, medications, procedure codes, diagnostic codes, text narrates from physicians and practitioners among patient's hospitalization. We selected patients who are associated

with both unstructured data and structured data. The statistics of the MIMIC-III dataset are summarized in Table 1.

Table 1. Statistics of MIMIC-III dataset

Characteristic	Number
Total patients	35,164
Single-visit patients	29,734
Multi-visit patients	5,159
Total diagnoses	6,646
Avg # of diagnoses	11.11
Total medications	155
Avg # of medication	9.23

In the pre-training phase, we used 80 percent single visit record (within both single-visit patients and multi-visit patients) as the training set. After obtaining the single-visit representation, we used the visit representation for the downstream tasks. For the drug recommendation task, we followed the G-BERT used the multi-visit sequence as input, we split the multi-visit patient records at the ratio of 0.85:0.5:0.1. For the 30-day readmission, we randomly selected 5,726 visit records and split the dataset with the ratio of 8:1:1. For the ICD coding task, we used the dataset analogous to the prior work CAML[22] with simple adaptations. The CAML dataset was preprocessed from the MIMIC-III dataset. To align with the input format for pre-training, we set the max token sequence length as 512. The dataset details of pre-training and fine-tuning are summarized in Table 2.

Table 2. Number of samples for pre-training, fine-tuning tasks

Task	Training set	Validating set	Testing set
Pre-training	39,550	-	-
Drug recommendation	4,344	272	543
ICD coding	8,066	1,573	1,729
30-day readmission	4,660	532	534

Model Overview

UMM-PLM handles the inputs from both unstructured and structured EHRs and fuses the information of two modalities in a unified manner. Figure 2 introduces the

architecture of UMM-PLM. Based on the input text sequence and code sequence, we designed a unimodal module, which learns unimodal data representations from the two modalities to preserve the modal-specific characteristics. More specifically, a BERT-like component and a G-BERT[47] like component are leveraged to encode the unstructured text and structured code inputs respectively. A multimodal cross-modal module is then introduced to integrate the interactive representation. Furthermore, specific pre-training tasks were designed to capture complementary information and model interactions between unstructured text and structured codes, i.e. the Text-to-Code prediction task and Code-to-Code prediction task.
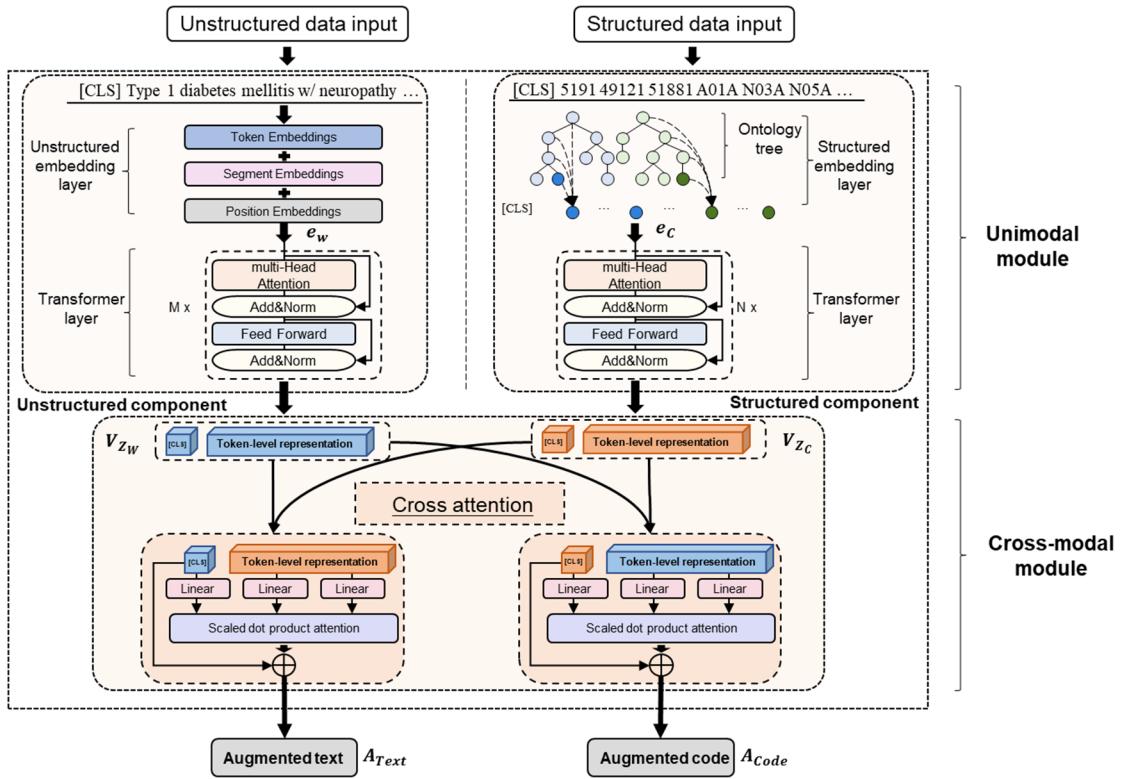

Figure 2. The model architecture of UMM-PLM. It includes a unimodal module that contains the unstructured text component and the structured code component (the upper part), and a cross-modal module (the bottom part).

Notions and Definitions

In the EHRs data, a patient's record is represented as a sequence of paired

unstructured data and structured data: $X = [(W_t, C_t)]$, where $t \in (1, 2, \ldots, T)$, T is the number of visits, where $W_t = (w_t^1, w_t^2, \cdots, w_t^k)$ is the word sequence of unstructured data records, and $C_t = (c_t^1, c_t^2, \cdots, c_t^l)$ is the corresponding structured data. Here we used $c \in C$ to represent either diagnoses code or medication codes. For each visit input, the unstructured data sequence is denoted as $W = ([CLS], w_1, w_2, \ldots, w_k)$ and the structured data can be represented as $C = ([CLS], c_1, c_2, \ldots, c_l)$, where $[CLS]$ is a special token to represent the summarized meaning of the overall sequence.

Unimodal Module

In the structured data embedding layer, inspired by G-BERT, [47] we used two ontologies, i.e. Anatomical Therapeutic Chemical, Third Level (ATC-3) and the International Classification of Diseases, Ninth Version (ICD-9) to categorize the medications and diagnosis codes, respectively, and each medical code is represented as a leaf node in the ontology tree. We firstly obtained the representation of each non-leaf node e_a in the ontology tree, where we applied Graph Attention Networks[56] to aggregate the representations of the non-leaf node and its direct children nodes to obtain the enhanced node representation e_a, and then fused the message passed from the ancestor nodes to each leaf node to get the enhanced embedding of the leaf node e_c.

$$e_a = GAT(A_{ai}, W_a) \tag{1}$$

$$e_c = GAT(A_{jc}, e_a) \tag{2}$$

Where W_a is code node initial embedding matrix, A_{ai} is the adjacent matrix of the non-leaf node n_a, and $i \in N_c$ is the direct child node of n_a. A_{jc} is the adjacent matrix of leaf node n_c, and $j \in N_a$ is the direct ancestor node. and GAT is the graph attention network.

After obtaining the enhanced embedding e_c of the leaf nodes, we used this enhanced feature to replace the randomly initialized embedding feature. Furthermore, unlike embeddings in the unstructured data, where position embeddings can be

leveraged to identify the relative token locations in context, a code in a visit usually does not have a fixed location assigned. Hence, position embeddings were not used.

In the unstructured embedding layer, the representation of the unstructured data is generated by summing the token embedding, segment embedding, and position embeddings similar to the original BERT[57] architecture.

$$e_w = SUM(e_{w_{token}}, e_{w_{segment}}, e_{w_{position}}) \tag{3}$$

Through the embedding layer of each, we obtained the text sequence representation e_w and code sequence representation e_c for each visit. Following previous work, we added a special token [CLS] as visit-level embedding to capture the patient feature in this visit. The multilayer Transformer[58] architecture was employed as the visit encoder for each modal. Each unimodal module takes its sequence representation as input and derives visit embeddings via the visit encoder.

$$Z_W = Transformer(e_w, \theta_w) \tag{4}$$

$$Z_C = Transformer(e_c, \theta_c) \tag{5}$$

where θ_w and θ_c are learnable parameters. Through the unimodal module, we obtained both the visit-level and token-level representations of each modality.

Cross-modal Module

To integrate the complementary information between data from multi-modalities, we designed a cross-attention mechanism to learn the interaction between the two modalities. For instance, we used the visit-level representation $v_{Z_{[CLS]}^C}$ from the structured data as a query to gather the complementary information from the unstructured data, then the token-level representations of the unstructured data as the key and value of the attention mechanism. Similar to the structured data, the unstructured data also performed the attention procedure to collect the structured information from the medical code sequence The cross-attention mechanism is defined

as follows:

$$A_{Text} = Softmax\left(\frac{Z_{W_{[CLS]}}Z_C^{\mathsf{T}}}{\sqrt{d_{Z_W}}}\right)Z_W + Z_{W_{[CLS]}} \tag{6}$$

$$A_{Code} = Softmax\left(\frac{Z_{C_{[CLS]}}Z_W^{\mathsf{T}}}{\sqrt{d_{Z_C}}}\right)Z_C + Z_{C_{[CLS]}} \tag{7}$$

where we use the scaled dot-product operation to measure the similarity between vectors. We also used a residual operator to augment unimodal-specific and cross-modal information. Our cross-modal module results in the augmented text representation A_{Text} and code representation A_{Code}.

Pre-training

Inspired by ALBEF[59], we modified the pre-training task Masked Language Model (MLM) into a multi-modal version. In particular, we defined two pre-training tasks, i.e. *Text-to-Code* and *Code-to-Code* (Figure 3). The modified pre-training tasks aim to learn the visit-level context-aware semantics of structured data from both the structured data sequence and unstructured data sequence. The pre-training objectives are:

$$\mathcal{L}_{T2C} = -\log\left(C_{[mask]}\middle|A_{Text_{(w_1,w_2,...,w_k)}}\right) \tag{8}$$

$$\mathcal{L}_{C2C} = -\log\left(C_{[mask]}\middle|A_{Code_{(c_1,c_2,...,c_l)}}\right) \tag{9}$$

where the $C_{[mask]}$ is the masked token from the input. L_{T2C} and L_{C2C} are cross-entropy loss functions.
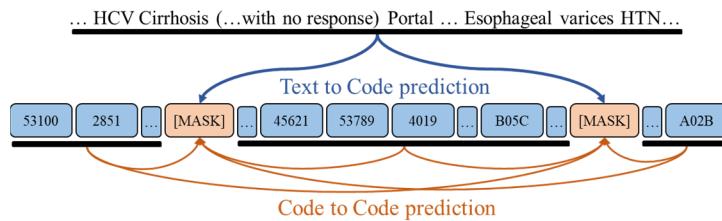

Figure 3. Graphical illustration of the pre-training tasks.

The final objective function is the sum of the above two losses:

$$\mathcal{L} = \mathcal{L}_{T2C} + \mathcal{L}_{C2C} \qquad (10)$$

In the unimodal module, for the pre-training of unstructured data, we initialized the model with pre-trained parameters from ClinicalBERT[46]. For the structured data, the parameters were randomly initialized. For the masked strategy of the two unimodal pre-training tasks, we randomly selected 15% tokens from the code sequence to mask. Rather than continually replacing the token chosen as the special token [MASK], the mask strategy is 80% rate to replace the token to [MASK] and 10% rate to keep the word not change, and 10% rate to replace to a random code.

Evaluation

After obtaining the pre-trained multimodal representations, UMM-PLM can be applied to downstream tasks through fine-tuning to improve the performance of these tasks. By adding task-adaptive classification layers, the downstream tasks could be binary classification (e.g., for patient 30-day readmission), multi-label classification (e.g., for drug recommendation), and else. We conducted experiments on three fine-tuning datasets extracted from MIMIC-III and a NER task on the n2c2 dataset.

Finetuning Task 1: Drug recommendation

Drug recommendation is an important application in healthcare, which aims to build computational models to automatically recommend medications (drugs) that are suited for a patient's health condition. This task is defined as utilizing the historical structured data records to predict the next-visit drug sequence. Similar to G-BERT, we utilized the diagnosis codes and medication codes from the historical visit records. Drug recommendation is a multi-label prediction task. For the code-aspect, we concatenated the mean of augmented representations of diagnoses and medications in the historical records (0 to $t-1$th visits) and the augmented representation of diagnoses for the t-

th visit, while for the text-aspect, we concatenated the mean of fused ***Text-to-Code*** representation of historical records (0 to $t-1$th visits) to predict the drug of the t-th visit. We built a multi-label Multi-Layer Perceptron (MLP) as the predication layer:

$$\mathcal{Y}_t^{drug} = \text{Sigmoid}\left(W^{drug}\left[E\left(\sum_{i<t} A_{Text_i}\right) \parallel \sum_{j=t} A_{Z_{C_d}}\right] \parallel E\left(\sum_{i<t-1} A_{Code_i}\right) + b^{drug}\right) \quad (11)$$

where $W^{drug} \in R^{3d_C+2d_W}$ is a learnable classification matrix, E is the expectation function, and "$\parallel$" is the concatenation operator. Given the ground truth labels $\hat{\mathcal{Y}}_t^{drug}$ of each timestamp, the loss function is:

$$\mathcal{L}_{drug} = -\frac{1}{T-1}\sum_{t=2}^{T}(\hat{\mathcal{Y}}_t^{drug}\log(\mathcal{Y}_t^{drug}) + (1-\hat{\mathcal{Y}}_t^{drug})\log(1-\mathcal{Y}_t^{drug}) \quad (12)$$

Finetuning Task 2: ICD coding

ICD coding for large-scale clinical notes is labor intensive and error prone, and machine learning methods could help reduce time and laborious cost in an automatic way[60]. ICD coding usually is treated as a multi-label classification problem. For the multi-modality input of EHRs, we also used the drug information in the corresponding visit as a complement of diagnosis coding. After the pre-training phase, we concatenated the outputs of the cross-modal module, which represents the visit-level text and code representations of the current record. Then an MLP classification layer is added to generate the ICD code:

$$\mathcal{Y}_{coding} = Sigmoid(W^{coding}[A_{Text}, A_{Code}] + b^{coding}) \quad (13)$$

where $W^{coding} \in R^{d_C+d_W}$ is the weight matrix of the output layer. The training objective is to minimize the binary cross-entropy loss between the prediction Y_{coding} and the target $\hat{\mathcal{Y}}_{coding}$:

$$\mathcal{L}_{coding} = -\sum_{j=1}^{J}\hat{\mathcal{Y}}_{coding}log(\mathcal{Y}_{coding}) + (1-\hat{\mathcal{Y}}_{coding})\log(1-\mathcal{Y}_{coding}) \quad (14)$$

Finetuning Task 3: 30-day readmission

The prediction of 30-day readmission is meaningful in practice in improving

patients' life quality and lower down the financial cost[61–63]. The task considers a patient

encounters a *"readmission"* if the admission date of patient was within 30 days after the

discharge date of the previous hospitalization, and thus is a binary classification task.

For each visit, we concatenated the pre-trained visit-level text feature A_{Z_W} and code

feature A_{Z_C} and used MLP to obtain the final output of the current visit:

$$\mathcal{Y}_{readm} = Sigmoid(W^{readm}[A_{Text}, A_{Code}] + b^{readm}) \tag{15}$$

$$\mathcal{L}_{readm} = -\sum_{p=1}^{P} \hat{\mathcal{Y}}_{readm} log(\mathcal{Y}_{readm}) + (1 - \hat{\mathcal{Y}}_{readm}) log(1 - \mathcal{Y}_{readm}) \tag{16}$$

where $\hat{\mathcal{Y}}_{readm}$ is the label of readmission, $W^{readm} \in R^{d_C + d_W}$ stands for the learnable

parameter.

Result

We compared UMM-PLM with several benchmark and state-of-the-arts methods, details in

Appendix A. Table 3 presents the results on the three downstream tasks, the best value

for each column bolded. The primary evaluation metric of the three tasks is Area Under

the Receiver Operating Characteristic (AUC). We also list the accuracies and F1s for

the first three tasks. From Table 3 we can generally draw the following conclusions: 1)

The deep learning-based methods perform much better than conventional machine

learning-based methods; 2) Methods with the addition of PLMs have better

performances than those without; 3) Combining structured and unstructured data do not

always obtain better results than using only a single-modal input; 4) Using UMM-PLM

obtains the best result in all tasks.

For example, in the drug recommendation task, G-BERT and Med-BERT

outperform LR and RNN when only code is taken as input. When combining code and

text, however, the direct concatenation methods, i.e. Med-BERT+ClinicalBERT and

G-BERT+ClinicalBERT performs worse than using G-BERT only. In comparison, our

model without modeling multimodal interactions (UMM-PLM$_{-cross\_modal}$) performs comparably with G-BERT. And when adding the cross-modal module, UMM-PLM improves G-BERT by 1.15%. In readmission prediction, using code only or text only can both achieve an AUC over 0.69, and ClinicalBERT+G-BERT with multimodal input can slightly outperform them. UMM-PLM-based models further improve the AUCs and UMM-PLM can even improve ClinicalBERT+G-BERT by 3.91%. Similar trends can also be observed in the task of ICD coding.

Table 3. Summary of performance on three downstream tasks on F1, Accuracy, and AUC. The standard deviations are listed in brackets.

Task	Model	F1%	Accuracy %	AUC%
Drug recommendation	LR(Code)	61.49(0)	89.07(0)	77.43(0)
	RNN[64] (Code)	58.48(0.05)	90.56(0.02)	91.95(0.04)
	G-BERT[65](Code)	65.75(0.33)	91.74(0.07)	94.40(0.06)
	Med-BERT[66](Code)	61.30(0.12)	90.93(0.01)	92.97(0.04)
	Med-BERT+ClinicalBERT	61.31(0.05)	90.91(0.02)	92.91(0.03)
	G-BERT+ClinicalBERT	65.35(0.21)	91.63(0.05)	94.30(0.06)
	UMM-PLM$_{-cross\_modal}$	66.10(0.18)	91.97(0.03)	94.39(0.02)
	UMM-PLM	**70.03(0.13)**	**92.82(0.05)**	**95.55(0.03)**
30-day readmission	CNN[67](Text)	57.84(1.70)	62.55(1.37)	66.74(1.02)
	ClinicalBERT[46](Text)	63.86(1.41)	64.19(1.40)	69.37(1.43)
	LR(Code)	63.17(0)	65.73(0)	65.73(0)
	G-BERT[65](Code)	64.82(1.02)	65.42(0.64)	69.57(0.43)
	Med-BERT[66](Code)	64.52(0.78)	64.22(0.97)	69.06(1.61)
	ClinicalBERT+G-BERT	65.63(1.12)	65.67(1.11)	70.79(0.33)
	ClinicalBERT+Med-BERT	64.35(0.85)	64.48(0.92)	69.38(1.26)
	UMM-PLM-$_{-cross\_modal}$	61.54(1.57)	66.54(0.87)	71.23(0.83)
	UMM-PLM	**68.61(0.83)**	**68.77(069)**	**74.70(0.50)**
ICD coding	CNN[67](Text)	49.08(0.73)	32.53(0.64)	84.06(0.68)
	ClinicalBERT[46](Text)	49.72(1.80)	33.10(1.59)	84.11(0.44)
	ClinicalBERT+G-BERT(Drug)	50.14(0.55)	33.46(0.49)	85.86(0.21)
	UMM-PLM$_{-cross\_modal}$	51.51(0.47)	34.07(0.03)	86.70(0.04)
	UMM-PLM	**52.09(0.65)**	**35.22(0.60)**	**87.46(0.05)**
*UMM-PLM-cross_modal means removing the multi-modal module from our model. The parentheses (Code) means using the structured data as input, and (Text) means using the unstructured data as input, and (Drug) means only use the drug code as the input of structured component.				

Further, in order to verify if UMM-PLM can be beneficial in different cases especially in scenarios with small training data, we conducted predictions on various training proportions by setting different training ratios. In Figure 4, the broken line graph shows the stability of the UMM-PLM model in outperforming other baselines. Even in the extreme circumstance where only 10% of the training set is available to train the downstream model, UMM-PLM still shows its superiority in contributing to all prediction tasks. In the ICD coding task in Figure 4(c), the pre-training models ClinicalBERT improve CNN by almost 5% when training on 10% of the train data, and UMM-PLM further improves ClinicalBERT by about 3%. We can also observe from Figure 4(a) and (c) that models without PLMs (e.g. RNN, CNN) have poor performances when the training size is extremely small (e.g. 10%).
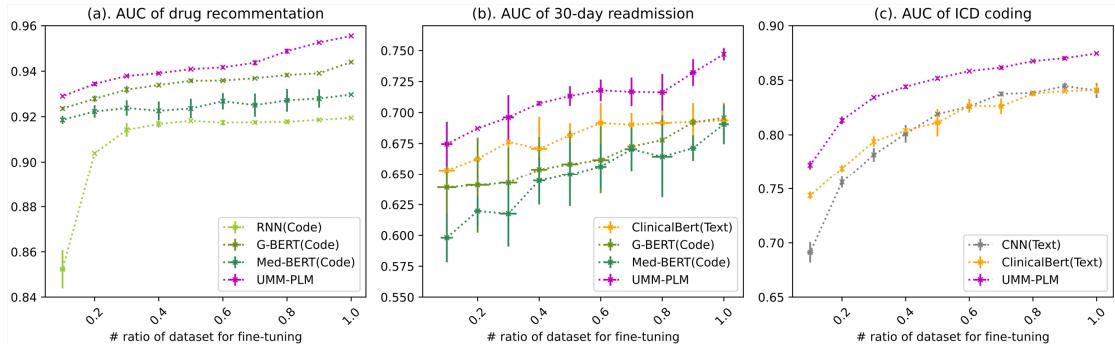

Figure 4. Comparisons of different models by using different training ratios.

Discussion

Upon analyzing the results in Table 3, we can conclude that the PLM-based methods obtain much better results than conventional machine learning and deep learning methods, e.g. LR and RNN, in general, and the results also verify that the modeling of multimodal data using our proposed UMM-PLM is effective across different tasks. Further experiments also demonstrate its stability on different sizes of training set on the fine-tuning phase. The success of UMM-PLM can be attributed to

that the pre-training phase well captures the complex and interactive semantics from multi-modal EHRs through two unsupervised pre-training tasks, Text-to-Code prediction and Code-to-Code prediction. Using Text-to-Code, the context of each code can be expanded by more detailed descriptions of the code or other correlated codes. For example, in Figure 1, the text in the purple box *Morphine 4mg* is an evidence to the coding of *N02C* (Migraine medication). Using Code-to-Code[2], the dependencies between different medical codes can be further enforced. We did not use Text-to-Text prediction since this process has already been well exploited in the pre-trained ClinicalBERT. The unsupervised pre-training tasks aim to re-construct the masked tokens using information from different modalities, thus can help the model collect intrinsic relationships among multimodal data, which is more powerful than direct concatenation, e.g. G-BERT+ClinicalBERT.

As mentioned above, using multimodal data do not always outperform unimodal methods. For example, in the drug recommendation task, the AUCs of G-BERT+ClinicalBERT, Med-BERT+ClinicalBERT and UMM-PLM-cross_modal are lower or only comparable with the unimodal pre-trained model G-BERT. This is partly determined by the characteristics of different scenarios. In drug recommendation, the basic predicted codes are inferences from the historical structured data, the diagnosis codes and historical medication codes reflected the health situation clearly. However, the unstructured data not only include related symptoms but also some extra information, including the family history and social history of patient, that impersonal information maybe disturbs the positive correlation of patient's further drug at the next visit.

---

[2] We re-added Code-to-Code prediction since the vocabulary of G-BERT is different from ours and we have to pre-train the parameters of the G-BERT component from scratch.

In Figure 4(c), the pre-trained model ClinicalBERT shows its ability in few-shot datasets, which benefit from its pre-trained parameters. However, as the dataset increases, the performance of CNN becomes comparable to pre-trained model, it may be due to the different assigned granularity that CNN used the fine-grained filter to assign candidate ICD code from sentence fragment, the pre-trained model ClinicalBERT used the special token [CLS] to predict code. Therefore, the benefits of better parameter initialization of pre-trained would gradually diminish with the growth of training samples. Compared with the ClinicalBERT, the special pre-train text-to-code task of UMM-PLM offered more code-related information.

To validate if multimodal information has been effectively utilized by UMM-PLM, we chose a case and visualizes its heat map generated by the attention weights of UMM-PLM, which shows the focus of the model by tagging the most informative words. According to Figure 5, the UMM-PLM can automatically assign variant weights to words in the unstructured data and codes in the structured data that have different importance in determination. These words and codes might be either corresponding, e.g. *edema* with *348.5* and *parasagittal meningioma* with *225.2*, or complementary, e.g. *Ativan IV* with *N05A* (antipsychotics) and *Fosphenytoin* with *N03A* (antiepileptics). Using this information, the EHR of a patient can be automatically tagged with cross-modal important signals identified.
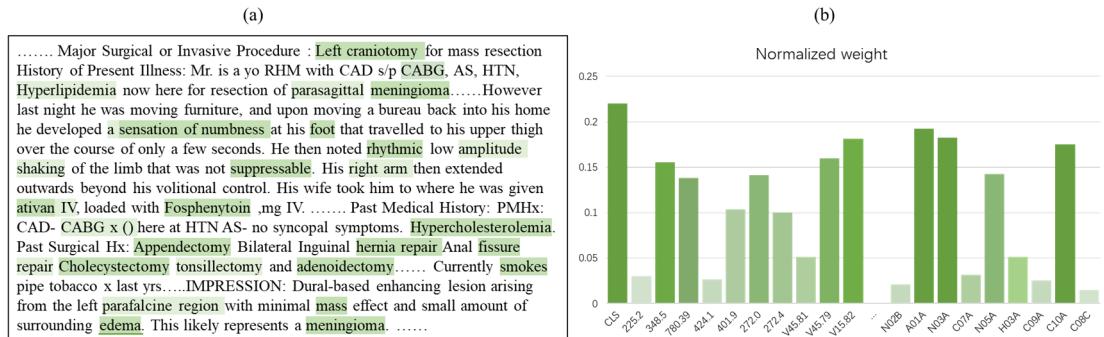

Figure 5. Visualization of attention weights for the text fragment and the code sequences. Where this patient's primary diagnosis is "BRAIN MASS". sub-figure (a) is the physician's narrative, and the (b) is

the corresponding structured code in EHRs. The gradient of the color indicates the importance.

In summary, the UMM-PLM model can efficiently integrate the interaction between structured data and unstructured data while maintaining the modal-specific representation capacity of the unimodal data. The pre-trained model has shown its robustness and potential value in clinical treatment, where solid performances on a variety of downstream tasks have been achieved. Furthermore, the model is also flexible according to its modularized structure that 1) It can be applied to cases where either multimodal or unimodal data are available, e.g. only using narratives for NER also works well; 2) The model can work in cases whether only medication or diagnostic codes are available. Just like the UMM-PLM in ICD coding task, where the UMM-PLM model only uses the structured drug sequence as input.

There are also several limitations of the current study. Firstly, we only selected medications and diagnoses for the structured data part and ignored others such as vital signs, laboratory measures, and procedure codes, which are also informative but might be complex to preprocess. Secondly, the UMM-PLM model only uses the single-visit record in the pre-training phase, and this strategy will ignore the time series and the continuity of the EHRs, which is also partly due to the nature of the MIMIC-III data. Thirdly, we truncated the length of each unstructured data sequence into a fixed number, which limited the content of the text information. Lastly, the MIMIC-III dataset only contains notes from the intensive care unit, and the performance of the model might be reduced on other clinical records. In future studies, we will explore adding more variables and the temporal patterns to our pre-trained model to improve the scalability and generalizability.

Conclusion

We propose a unified medical multimodal pre-training model named UMM-PLM in this work. The model was pre-trained to capture both unimodal representation abilities and cross-model interactions from EHRs and was evaluated on three downstream tasks. Experiments demonstrate the superiority and stability of the UMM-PLM model. We also tested the performance of UMM-PLM on smaller training sets, which further verified the capability of the model in few-shot learning cases. We expect our model could assist in more application scenarios where both structured and unstructured EHRs are available.

Data Availability

The pre-training dataset Medical Information Mart for Intensive Care III (MIMIC-III 1.4) is publicly available from the https://mimic.physionet.org/. This is a restricted-accessed resource hosted by PhysioNet, which should be used under license from the

credentialed account of PhysioNet. The downstream tasks of drug recommendations, ICD coding, and 30-day readmission are also associated with MIMIC-III. The dataset for the NER task is open-resource and available at https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/.

Code Availability

The codes of the UMM-PLM model and the pre-trained model are shared here: https://git.openi.org.cn/liusc/3-6-liusicen-multi-modal-pretrain.

Disclaimer

The content is solely the responsibility of the authors and does not necessarily represent the official views of the Cancer Prevention and Research Institute of Texas. The authors have no competing interests to declare.

References

1. Jensen, P. B., Jensen, L. J. & Brunak, S. Mining electronic health records: Towards better research applications and clinical care. *Nature Reviews Genetics* **13**, 395–405 (2012).

2. Rajkomar, A. *et al.* Scalable and accurate deep learning with electronic health records. *npj Digit. Med.* **1**, 18 (2018).

3. Lee, C. *et al.* Big Healthcare Data Analytics: Challenges and Applications. in 11–41 (Springer, Cham, 2017). doi:10.1007/978-3-319-58280-1_2

4. Murdoch, T. B. & Detsky, A. S. The inevitable application of big data to health care. *JAMA - Journal of the American Medical Association* **309**, 1351–1352 (2013).

5. Shickel, B., Tighe, P. J., Bihorac, A. & Rashidi, P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE J. Biomed. Heal. Informatics* **22**, 1589–1604 (2018).

6. Marafino, B. J. *et al.* Validation of Prediction Models for Critical Care Outcomes Using Natural Language Processing of Electronic Health Record Data. *JAMA Netw. Open* **1**, e185097–e185097 (2018).

7. Moskovitch, R., Polubriaginof, F., Weiss, A., Ryan, P. & Tatonetti, N. Procedure prediction from symbolic Electronic Health Records via time intervals analytics. *J. Biomed. Inform.* **75**, 70–82 (2017).

8. Beam, A. L. *et al.* Clinical concept embeddings learned from massive sources of multimodal medical data. in *Pacific Symposium on Biocomputing* **25**, 295–306 (World Scientific Publishing Co. Pte Ltd, 2020).

9. Gottesman, O. *et al.* The Electronic Medical Records and Genomics (eMERGE) Network: Past, present, and future. *Genetics in Medicine* **15**, 761–771 (2013).

10. Kawamoto, K., Houlihan, C. A., Balas, E. A. & Lobach, D. F. Improving clinical practice using clinical decision support systems: A systematic review of trials to identify features critical to success. *British Medical Journal* **330**, 765–768 (2005).

11. Carey, D. J. *et al.* The Geisinger MyCode community health initiative: An electronic health record-linked biobank for precision medicine research. *Genet. Med.* **18**, 906–913 (2016).

12. Andreu-Perez, J., Poon, C. C. Y., Merrifield, R. D., Wong, S. T. C. & Yang, G. Z. Big Data for Health. *IEEE J. Biomed. Heal. Informatics* **19**, 1193–1208 (2015).

13. Venugopalan, J., Tong, L., Hassanzadeh, H. R. & Wang, M. D. Multimodal deep learning models for early detection of Alzheimer's disease stage. *Sci. Rep.* **11**, 1–13 (2021).

14. Tan, Q. *et al.* DATA-GRU: Dual-attention time-aware gated recurrent unit for irregular multivariate time series. *AAAI 2020 - 34th AAAI Conf. Artif. Intell.* 930–937 (2020). doi:10.1609/aaai.v34i01.5440

15. Kwak, H., Chang, J., Choe, B., Park, S. & Jung, K. Interpretable disease prediction using heterogeneous patient records with self-attentive fusion encoder. *J. Am. Med. Informatics Assoc.* **28**, 2155–2164

16. Duan, H., Sun, Z., Dong, W., He, K. & Huang, Z. On Clinical Event Prediction in Patient Treatment Trajectory Using Longitudinal Electronic Health Records. *IEEE J. Biomed. Heal. Informatics* **24**, 2053–2063 (2020).

17. Tsai, S.-C., Huang, C.-W. & Chen, Y.-N. Modeling Diagnostic Label Correlation for Automatic ICD Coding. 4043–4052 (2021). doi:10.18653/v1/2021.naacl-main.318

18. McDermott, M. *et al.* A Comprehensive EHR Timeseries Pre-training Benchmark ACM Reference Format. *ACM CHIL '21*,

19. Ma, F. *et al.* A General Framework for Diagnosis Prediction via Incorporating Medical Code Descriptions. *Proc. - 2018 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2018* 1070–1075 (2019). doi:10.1109/BIBM.2018.8621395

20. Cheng, Y., Wang, F., Zhang, P. & Hu, J. Risk prediction with electronic health records: A deep learning approach. in *16th SIAM International Conference on Data*

Mining 2016, SDM 2016 432–440 (Society for Industrial and Applied Mathematics Publications, 2016). doi:10.1137/1.9781611974348.49

21. Choi, E. *et al.* Learning the Graphical Structure of Electronic Health Records with Graph Convolutional Transformer. *Proc. AAAI Conf. Artif. Intell.* **34**, 606–613 (2020).
22. Jensen, P. B., Jensen, L. J. & Brunak, S. Mining electronic health records: Towards better research applications and clinical care. *Nature Reviews Genetics* **13**, 395–405 (2012).
23. Miotto, R., Li, L., Kidd, B. A. & Dudley, J. T. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Sci. Rep.* **6**, 1–10 (2016).
24. Ching, T. *et al.* Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **15**, (2018).
25. Deng, Y. *et al.* A multimodal deep learning framework for predicting drug-drug interaction events. *Bioinformatics* **36**, 4316–4322 (2020).
26. Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A. & Escobar, G. Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Aff.* **33**, 1123–1131 (2014).
27. Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G. & Galstyan, A. Multitask learning and benchmarking with clinical time series data. *Sci. Data* **6**, 1–18 (2019).
28. Zheng, Z., Wang, Y., Dai, Q., Zheng, H. & Wang, D. Metadata-driven Task Relation Discovery for Multi-task Learning. 4426–4432 (2019). doi:10.24963/ijcai.2019/615
29. Johansen, M. L. & O'Brien, J. L. Decision Making in Nursing Practice: A Concept Analysis. *Nurs. Forum* **51**, 40–48 (2016).
30. Tiffen, J., Corbridge, S. J. & Slimmer, L. Enhancing Clinical Decision Making: Development of a Contiguous Definition and Conceptual Framework. *J. Prof. Nurs.* **30**, 399–405 (2014).
31. Qiao, Z., Wu, X., Ge, S. & Fan, W. MNN: Multimodal attentional neural networks for diagnosis prediction. *IJCAI Int. Jt. Conf. Artif. Intell.* **2019**-**Augus**, 5937–5943 (2019).
32. Xu, Z., So, D. R. & Dai, A. M. MUFASA: Multimodal Fusion Architecture Search for Electronic Health Records. (2021).
33. Iakovidis, D. & Smailis, C. A semantic model for multimodal data mining in healthcare information systems. in *Studies in Health Technology and Informatics* **180**, 574–578 (IOS Press, 2012).
34. Pivovarov, R. *et al.* Learning probabilistic phenotypes from heterogeneous EHR data. *J. Biomed. Inform.* **58**, 156–165 (2015).
35. Chen, J. H., Goldstein, M. K., Asch, S. M., Mackey, L. & Altman, R. B. Predicting inpatient clinical order patterns with probabilistic topic models vs conventional order sets. *J. Am. Med. Informatics Assoc.* **24**, 472–480 (2017).
36. Si, Y., Wang, J., Xu, H. & Roberts, K. Enhancing clinical concept extraction with contextual embeddings. *J. Am. Med. Informatics Assoc.* **26**, 1297–1304 (2019).
37. Khattak, F. K. *et al.* A survey of word embeddings for clinical text. *Journal of Biomedical Informatics: X* **4**, 100057 (2019).
38. Li, H. *et al.* CNN-based ranking for biomedical entity normalization. *BMC Bioinformatics* **18**, 280–286 (2017).
39. Wu, S. *et al.* Deep learning in clinical natural language processing: A methodical review. *Journal of the American Medical Informatics Association* **27**, 457–470 (2020).
40. Lin, C., Miller, T., Dligach, D., Bethard, S. & Savova, G. A BERT-based Universal Model for Both Within- and Cross-sentence Clinical Temporal Relation Extraction. *Proc. 2nd Clin. Nat. Lang. Process. Work.* **2**, 65–71 (2019).
41. Alsentzer, E. *et al.* Publicly Available Clinical BERT Embeddings. *Proc. 2nd Clin. Nat. Lang. Process. Work.* 72–78 (2019).
42. Lee, J. *et al.* BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
43. Ji, Z., Wei, Q. & Xu, H. BERT-based Ranking for Biomedical Entity Normalization.

AMIA Summits Transl. Sci. Proc. **2020**, 269 (2020).

44.   van Aken, B. *et al.* Clinical outcome prediction from admission notes using self-supervised knowledge integration. in *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference* 881–893 (2021). doi:10.18653/v1/2021.eacl-main.75

45.   Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.* **1**, 4171–4186 (2019).

46.   Huang, K., Altosaar, J. & Ranganath, R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. (2019).

47.   Shang, J., Ma, T., Xiao, C. & Sun, J. Pre-training of graph augmented transformers for medication recommendation. *IJCAI Int. Jt. Conf. Artif. Intell.* **2019-Augus**, 5953–5959 (2019).

48.   Li, Y. *et al.* BEHRT: Transformer for Electronic Health Records. *Sci. Rep.* **10**, 1–12 (2020).

49.   Rasmy, L., Xiang, Y., Xie, Z., Tao, C. & Zhi, D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digit. Med.* **4**, (2021).

50.   Antol, S. *et al.* VQA: Visual question answering. in *Proceedings of the IEEE International Conference on Computer Vision* **2015 Inter**, 2425–2433 (2015).

51.   Wang, Q., Zhan, L., Thompson, P. & Zhou, J. Multimodal Learning with Incomplete Modalities by Knowledge Distillation. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* 1828–1838 (2020). doi:10.1145/3394486.3403234

52.   Xu, Y., Biswal, S., Deshpande, S. R., Maher, K. O. & Sun, J. RAIM: Recurrent attentive and intensive model of multimodal patient monitoring data. in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2565–2573 (2018). doi:10.1145/3219819.3220051

53.   Gao, J., Li, P., Chen, Z. & Zhang, J. A survey on deep learning for multimodal data fusion. *Neural Computation* **32**, 829–864 (2020).

54.   Liu, J., Hai, Z., Yang, M. & Bing, L. Multi-perspective coherent reasoning for helpfulness prediction of multimodal reviews. in *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference* 5927–5936 (2021). doi:10.18653/v1/2021.acl-long.461

55.   Ngiam, J. *et al.* Multimodal deep learning. in *Proceedings of the 28th International Conference on Machine Learning, ICML 2011* 689–696 (2011).

56.   Veličković, P. *et al.* Graph attention networks. in *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings* (2018).

57.   Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.* **1**, 4171–4186 (2018).

58.   Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (2018).

59.   Li, J. *et al.* Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. *arxiv.org* (2021).

60.   Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J. & Eisenstein, J. Explainable prediction of medical codes from clinical text. *NAACL HLT 2018 - 2018 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.* **1**, 1101–1111 (2018).

61.   Kansagara, D. *et al.* Risk prediction models for hospital readmission: A systematic review. *JAMA - Journal of the American Medical Association* **306**, 1688–1698 (2011).

62.   Kripalani, S., Theobald, C. N., Anctil, B. & Vasilevskis, E. E. Reducing hospital readmission rates: Current strategies and future directions. *Annual Review of Medicine*

65, 471–485 (2014).

63.   Felix, H. C., Seaberg, B., Bursac, Z., Thostenson, J. & Stewart, M. K. Why Do Patients Keep Coming Back? Results of a Readmitted Patient Survey. *Soc. Work Health Care* **54**, 1–15 (2015).

64.   Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).

65.   Shang, J., Ma, T., Xiao, C. & Sun, J. Pre-training of graph augmented transformers for medication recommendation. in *IJCAI International Joint Conference on Artificial Intelligence* **2019**-**Augus**, 5953–5959 (2019).

66.   Rasmy, L., Xiang, Y., Xie, Z., Tao, C. & Zhi, D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digit. Med.* **4**, (2021).

67.   Dos Santos, C. N. & Gatti, M. Deep convolutional neural networks for sentiment analysis of short texts. in *COLING 2014 - 25th International Conference on Computational Linguistics, Proceedings of COLING 2014: Technical Papers* 69–78 (2014).

68.   Paszke, A. *et al. Automatic differentiation in pytorch*. (2017).

69.   Cho, K. *et al.* Learning phrase representations using RNN encoder-decoder for statistical machine translation. in *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* 1724–1734 (2014). doi:10.3115/v1/d14-1179

70.   Kim, Y. Convolutional neural networks for sentence classification. in *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* 1746–1751 (2014). doi:10.3115/v1/d14-1181

71.   Uzuner, Ö., South, B. R., Shen, S. & DuVall, S. L. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J. Am. Med. Informatics Assoc.* **18**, 552–556 (2011).

72.   Chalapathy, R., Borzeshi, E. Z. & Piccardi, M. Bidirectional LSTM-CRF for Clinical Concept Extraction. (2016).

73.   Li, J. *et al.* Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. 1–16 (2021).

74.   He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning. in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 9726–9735 (2020). doi:10.1109/CVPR42600.2020.00975

Appendix

A.   *Baselines*

We compared UMM-PLM with the following baselines. All methods are implemented in Pytorch[68].

LR: Logistic Regression was a convention machine learning method, it was explored for ICD coding task by building binary one-versus-rest classifiers, and was explored for 30-day readmission task by building binary classifies.

RNN: Recurrent Neural Network[69] used the patient sequence as input to learn the hidden representation of patient, and convert the hidden state as a binary classification.

CNN: The one-dimensional Convolutional Neural Network[70] was employed for ICD coding task, which treated this automatic coding as a multi-label classification task[60].

G-BERT: The G-BERT proposed combined Graph Neural Networks and BERT for medical code representation, which use GNNs to represent the hierarchical structures of medical code, and integrate the GNNs representation into transformer-based pre-train model[47].

Med-BERT: Med-BERT adapts the BERT framework for the text domain to the structured EHRs, which defined serialization embeddings to denote the relative order of each code[49].

ClinicalBERT: The ClinicalBERT pre-train BERT using the clinical notes and fine-tune the pre-trained network for the task of predicting hospital readmission[46].

G-BERT+ClinicalBERT: We concatenated the structured code representation of G-BERT and unstructured data representation of ClinicalBERT straightforwardly. This method was used to verify the different combined ways of multimodal.

Med-BERT+ClinicalBERT: Similar to the G-BERT+ClinicalBERT, the Med-

BERT+ClinicalBERT concatenated the pre-trained representation of Med-BERT and ClinicalBERT.

UMM-PLM-cross_modal: This is a variant of UMM-PLM, UMM-PLM-cross_modal removes the cross-modal module to conduct the ablation experiments.

B.   *Implementation Details*

In the unimodal module, for the unstructured data component, we used 12 encoder layers, 12 attention heads, and a hidden dimension of $768(L = 12, H = 768, A = 12)$. For the structured data component, we used 2 encoder layers, 2 attention heads, and a hidden size of 300. The ontology embedding size is 75, and the number of heads for ontology aggregation attention is 4.

We used the PyTorch BERT codebase for implementation and set the maximum sequence length of unstructured data as 512 tokens. Since we did not use the Text-to-Text MLM, to better inherit the pre-trained parameters from ClinicalBERT and align with the structured data component, we froze the previous ten layers of the encoder and the embedding layer of ClinicalBERT and only optimized the parameters of the last two encoder layers. The maximum sequence length for the structured data was set as 61, which is the maximum number of codes in a single visit. We masked the structured data using a 15% rate similar to the original BERT. We used the learning rate of $5e - 4$ and dropout rate of 0.1, and training batch size of 32. Training is done through the Adam[24] optimizer. The model was trained on the evaluation set with a maximum of 200 epochs. Two GeForce RTX 3090 GPUs were leveraged to pre-train the UMM-PLM model and the early-stopping method was also utilized.

In the fine-tuning phase, we set different learning rates for the tasks, e.g. $5e - 5$ for drug recommendations, $2e - 5$ for 30-day readmission, $1e - 5$ for ICD coding, and $3e - 5$ for NER. All evaluations were duplicated five times with different random

seeds to eliminate overfitting and the average values and standard deviations of evaluation metrics were reported.

C.   Finetuning Task: Named Entity Recognition (NER)

Biomedical Named Entity Recognition (NER) was also taken as an auxiliary task to validate if the multimodal pre-trained model could also perform well on a single-modal task. We adopted the concept extraction task of the 2010 i2b2 2/VA Workshop on Natural Language Processing Challenges for Clinical Records[71] for the NER task. the dataset details are shown in Table 4:

Table 4 Details of 2010 i2b2 NER dataset

	NER (2010-i2b2)		
Dataset	Training set	Validing set	Testing set
Sentence	14,803	1,512	27,625

The NER task focuses on extracting medical concepts from patient reports, which is purely based on unstructured data. We performed the NER task based on the BiLSTM-CRF[72] framework and replaced the embedding layer of the original BiLSTM-CRF with the output of the unstructured data module, and for the NER task, the primary metric is strict F1 score.

Table 5 Performance of NER task

Task	Model	F1%
NER	biLSTM_CRF[72]	83.81
	biLSTM_CRF$_{+ClinicalBert}$	85.77
	biLSTM_CRF$_{+UMM-PLM}$	**86.29**

AS shown in Table 5, for NER, comparing the results of biLSTM_CRF$_{+UMM-PLM}$ and biLSTM_CRF$_{+ClinicalBERT}$, we have verified that UMM-PLM still maintains a good representation capacity for free text.

D.   Ablation experiment

We replaced the pre-trained model parameters of G-BERT from UMM-PLM to verified the special pre-train task was benefited the structured data. And the results are shown

in Tabel 6:

Table 6. Performance of Drug recommendation task of ablation experiment

Task	Model	F1%	Accuracy%	AUC%
Drug recommendation	G-BERT[65]	65.75(0.33)	91.74(0.07)	94.40(0.06)
	G-BERT(UMM-PLM)	67.84(0.12)	92.23(0.04)	95.15(0.02)

E.   Contrastive learning of UMM-PLM

Contrastive self-supervised learning techniques are a promising class of methods that build representations by learning to maximize the different data and minimum similar data[73]. We explore the Contrastive Learning (CL) method on the UMM-PLM model. In the pre-training phase, we add a Text-Code contrastive loss to model the text and code modality pair. The contrastive loss aims to make Text-Code pairs semantic distance closer in the same patient and farther among different patients. We use the momentum contrast method[74] to build a momentum module. The momentum encodes the similarity of Text-Code pairs as the pseudo label. Then compute the contrastive loss between the soft pseudo label and the pre-training similarity label. For each Text-Code pair, we calculate the text-to-code and code-to-text similarity as:

$$S(A,B) = MLP_1(A)^\mathrm{T} MLP_2(B) \tag{1}$$

$$y_{t2c} = \frac{\exp{(S(X_w, X_c)/\tau)}}{\sum_{m=1}^{M} \exp{(S(X_w, X_c)/\tau)}} \tag{2}$$

$$y_{c2t} = \frac{\exp{(S(X_c, X_w)/\tau)}}{\sum_{m=1}^{M} \exp{(S(X_c, X_w)/\tau)}} \tag{3}$$

$$y_{t2c}^m = MoCo_{t2c}(X_w, X_c) \tag{4}$$

$$y_{c2t}^m = MoCo_{c2t}(X_w, X_c) \tag{5}$$

$$\mathcal{L}_{cl} = (1-\alpha)y_{pairs} + \alpha[KL(y_{t2c}^m||y_{t2c}), KL(y_{c2t}^m||y_{c2t})] \tag{6}$$

where the y_{pairs} are the positive pairs that the unstructured data and structured data pair is the same patient, $y_{pairs} \in [0,1]$, y_{t2c} and y_{c2t} is the similarity score of the unstructured data and structured data. The results of the drug recommendations task

using the contrastive learning method are shown in Table 7.

Table 7 performance of Contrastive Learning method on drug recommendations task

Task	Model	F1%	Accuracy%	AUC%
Drug recommendations	UMM-PLM	70.03(0.13)	92.82(0.05)	95.55(0.03)
	UMM-PLM+CL	71.95(0.61)	93.33(0.13)	95.76(0.03)