frontiers in Public Health

ORIGINAL RESEARCH
published: 20 January 2022
doi: 10.3389/fpubh.2021.793801

Unifying Diagnosis Identification and Prediction Method Embedding the Disease Ontology Structure From Electronic Medical Records

Jingfeng Chen[1,2]*, *Chonghui Guo*[2]*, *Menglin Lu*[2] *and Suying Ding*[1]

[1] Health Management Center, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China, [2] School of Economics and Management, Institute of Systems Engineering, Dalian University of Technology, Dalian, China

Objective: The reasonable classification of a large number of distinct diagnosis codes can clarify patient diagnostic information and help clinicians to improve their ability to assign and target treatment for primary diseases. Our objective is to identify and predict a unifying diagnosis (UD) from electronic medical records (EMRs).

Methods: We screened 4,418 sepsis patients from a public MIMIC-III database and extracted their diagnostic information for UD identification, their demographic information, laboratory examination information, chief complaint, and history of present illness information for UD prediction. We proposed a data-driven UD identification and prediction method (UDIPM) embedding the disease ontology structure. First, we designed a set similarity measure method embedding the disease ontology structure to generate a patient similarity matrix. Second, we applied affinity propagation clustering to divide patients into different clusters, and extracted a typical diagnosis code co-occurrence pattern from each cluster. Furthermore, we identified a UD by fusing visual analysis and a conditional co-occurrence matrix. Finally, we trained five classifiers in combination with feature fusion and feature selection method to unify the diagnosis prediction.

Results: The experimental results on a public electronic medical record dataset showed that the UDIPM could extracted a typical diagnosis code co-occurrence pattern effectively, identified and predicted a UD based on patients' diagnostic and admission information, and outperformed other fusion methods overall.

Conclusions: The accurate identification and prediction of the UD from a large number of distinct diagnosis codes and multi-source heterogeneous patient admission information in EMRs can provide a data-driven approach to assist better coding integration of diagnosis.

Keywords: unifying diagnosis, disease ontology structure, set similarity measure, clustering, electronic medical records

OPEN ACCESS

Edited by:
Yi-Ju Tseng,
National Central University, Taiwan

Reviewed by:
Martin Hofmann-Apitius,
Fraunhofer Institute for Algorithms and Scientific Computing (FHG), Germany
Hsin-Yao Wang,
Linkou Chang Gung Memorial Hospital, Taiwan

***Correspondence:**
Chonghui Guo
dlutguo@dlut.edu.cn
Jingfeng Chen
fccjfchen@zzu.edu.cn

Specialty section:
This article was submitted to Digital Public Health, a section of the journal Frontiers in Public Health

Received: *12 October 2021*
Accepted: *21 December 2021*
Published: *20 January 2022*

Citation:
Chen J, Guo C, Lu M and Ding S (2022) Unifying Diagnosis Identification and Prediction Method Embedding the Disease Ontology Structure From Electronic Medical Records. Front. Public Health 9:793801. doi: 10.3389/fpubh.2021.793801

INTRODUCTION

In medical practice, clinicians are encouraged to seek a unifying diagnosis (UD) that could explain all the patient's signs and symptoms in preference to providing several explanations for the distress being presented (1). A UD is a critical pathway to identify the correct illness and craft a treatment plan; thus, clinical experience and knowledge play an important role in the science of diagnostic reasoning. Generally, from a brief medical history from a patient, clinicians can use the intuitive system in their brain and rapidly reason the disease types, whereas for complex and multi-type abnormal results, clinicians must use the more deliberate and time-consuming method of analytic reasoning to deduce the UD, raising the risk of diagnostic errors (2).

To increase the accuracy of a UD, enhancing individual clinicians' diagnostic reasoning skills and improving health care systems are regarded as two important approaches to support clinicians through the diagnostic process. The former requires professional knowledge training and lifelong learning, whereas the latter mainly involves the development of information technology (3). For an individual clinician, an intelligent clinical decision support system is prone to acceptable and can help clinicians to improve their unifying diagnostic decisions (4). Recently, along with the widespread adoption of electronic medical records (EMRs), an extremely large volume of electronic clinical data has been generated and accumulated (5, 6). Meanwhile, artificial intelligence and big data analytic technology have been successfully applied to clinical diagnostic procedures and treatment regimen recommendation, which has resulted in new opportunities for intelligent clinical decision support systems that use data-driven knowledge discovery methods (7–10).

From the data mining perspective, a UD aims to classify a large number of distinct diagnosis codes reasonably according to the disease taxonomy and attempt to adopt a disease to summarize or explain various clinical manifestations of the disease. Therefore, the nature of a UD is diagnosis code assignment along with disease correlation exploitation. Diagnosis code assignment refers to the clinical decision process in which supervised methods are adopted to predict and annotate disease codes based on patients' medical history, signs and symptoms, and laboratory examination (11). According to the number of diagnosis codes that patients suffer from, diagnosis code assignment can be divided into single-label (12), multi-class (13), multi-label (14), and multi-task learning methods (15). However, although many novel supervised learning models have been proposed and can achieve high performance in terms of assigning diagnosis codes for new patients using frontier supervised methods, such as ensemble learning (16), reinforcement learning (17), and deep learning (18), they cannot further explore disease co-occurrence relations for UD identification and prediction.

The coexistence of multiple diseases is pervasive in the clinical environment, particularly for patients in the intensive care unit (ICU) (19). According to the statistical results of the MIMIC-III database, which is a freely accessible critical care database, the average number of diagnosis codes for patients in the ICU is 11. Additionally, diagnosis codes are highly fine-grained, closely related, and extremely diverse (20). For example, the patient with admission identifier (ID) 100223 is assigned to 28 ICD-9 codes, and many diagnosis codes are similar, such as 276.2 (Acidosis, order: 15), 276.0 (Hyperosmolality and/or hypernatremia, order: 18), and 276.6 (Hyperpotassemia, order: 26). Thus, it is trivial and difficult for clinicians to make a consistent, accurate, concise, and unambiguous diagnostic decision reasonably.

Furthermore, although the inter-relation of diagnosis codes was considered in previous studies, the researchers commonly used the first three digits of ICD-9 codes to assign diagnosis codes for patients (21–23); hence, the complexity may increase and prediction performance may reduce when considering all digits of the ICD-9 codes. Additionally, in those studies, reasonable complicated and confused diagnosis codes could not be classified into a UD using a data-driven method. A UD is the basic principle of clinical diagnostic thinking. Its basic idea is that when a patient has many symptoms, if these symptoms can be explained by one disease, it will never explain different symptoms using multiple diseases (1). A UD reflects the integrity of the patient and the professionalism of clinicians; however, in previous studies, the main focus was on the UD of a category of diseases from the clinical perspective, such as mood/mental disorders (24), intracranial mesenchymal tumor (25), and arrhythmogenic right ventricular cardiomyopathy (26). In this study, we fully consider the fine-grained diagnosis codes (i.e., all digits) of patients, identify the UD from a group of patient diagnostic information using an unsupervised clustering method and predict the UD for new unseen patients using multi-class learning methods.

MATERIALS AND METHODS

Data Collection

We selected a dataset of sepsis patients from the MIMIC-III database, where sepsis is divided into general sepsis, severe sepsis, and septic shock (27, 28). **Figure 1** shows the detailed processes of data collection and preprocessing of sepsis patients, including the identification of sepsis patients, data extraction, data cleaning, and feature selection. Finally, we screened 4,418 sepsis patients and extracted their diagnostic information to unify the diagnosis identification, their demographic information, laboratory examination information, chief complaint, and history of present illness information, and obtain a UD prediction.

First, the diagnostic information of 4,418 sepsis patients mainly contained the patient hospital admission ID (Hadm-id), ICD-9 diagnosis code, order of diagnosis code, and a brief definition of the diagnosis codes, where the sum, maximum, minimum, and average numbers of diagnosis codes were 80501, 39, 3, and 18.3, respectively. Additionally, for the visualization,

Abbreviations: EMR, Electronic medical record; UDIPM, Unifying diagnosis identification and prediction method; CDSS, Clinical decision support system; ICU, Intensive care unit; IC, Information content; LCA, Least common ancestor; AP, Affinity propagation; SS, Sum of similarities; TDC, Typical diagnosis code; LCoP, LCA co-occurrence pattern; AOrd, Average order; TDCCoP, Typical diagnosis code co-occurrence pattern; CCoM, Conditional co-occurrence matrix; UD, Unifying diagnosis; Hadm-id, Hospital admission identifier; FM, Fusion method.

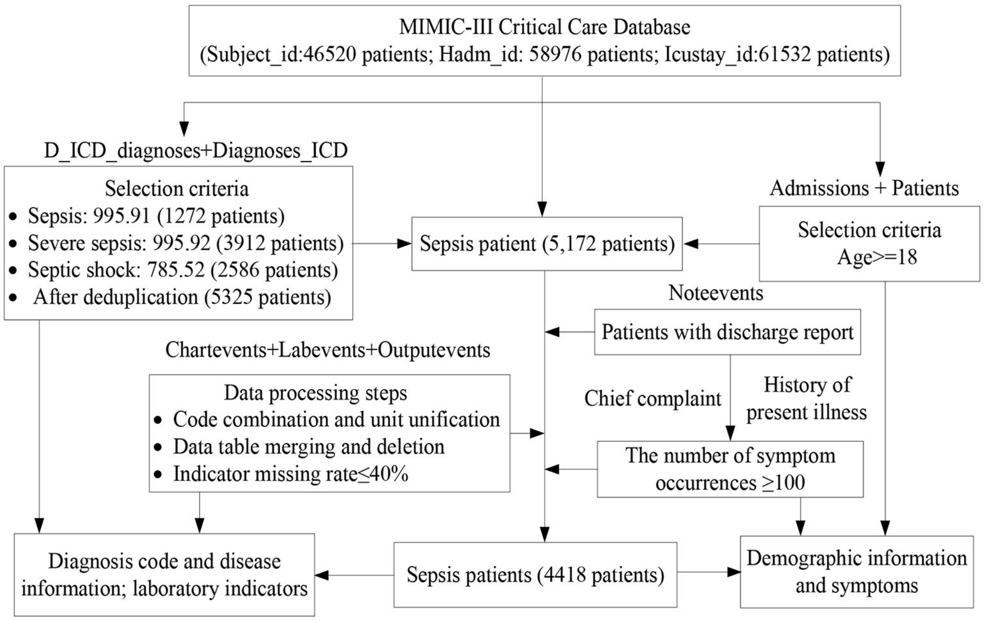
MIMIC-III Critical Care Database
(Subject_id:46520 patients; Hadm_id: 58976 patients; Icustay_id:61532 patients)
D_ICD_diagnoses+Diagnoses_ICD
Selection criteria
• Sepsis: 995.91 (1272 patients)
• Severe sepsis: 995.92 (3912 patients)
• Septic shock: 785.52 (2586 patients)
• After deduplication (5325 patients)
Sepsis patient (5,172 patients)
Admissions + Patients
Selection criteria
Age>=18
Noteevents
Patients with discharge report
Chartevents+Labevents+Outputevents
Data processing steps
• Code combination and unit unification
• Data table merging and deletion
• Indicator missing rate≤40%
Chief complaint
History of present illness
The number of symptom occurrences ≥100
Diagnosis code and disease information; laboratory indicators
Sepsis patients (4418 patients)
Demographic information and symptoms

FIGURE 1 | Dataset selection of sepsis patients from the MIMIC-III database.

TABLE 1 | Feature information of the health condition of sepsis patients.

Information	Feature	Description (Range, Type)
Demographic information	Admission type	Emergency, elective, urgent (Nominal)
	Gender	Female, male (Nominal)
	Age	[18, 89] (Numeric)
Laboratory examination information	Potassium Level, PO2, serum bicarbonate level, temperature, sodium level, urine out foley, urea nitrogen, WBC, bilirubin level, GCSmotor, GCSeyes, HR, GCSverbal, NBP, RR, SPO2, hemoglobin, platelet count, creatimine	Minimum, maximum, median, mean, and variance value (Numeric)
Symptom information	Fever, abdominal pain, shortness of breath, nausea and vomiting, weakness, diarrhea, dizziness, palpitation, cough, fatigue, discomfort, dysuria, shock, weight change, loss of appetite, and night sweating	0, 1 (Nominal)
Related indicators	AIDS, hematologic malignancy, metastatic cancer	0, 1 (Nominal)
	SOFA, SAPS, and SAPS-II	Integer (Numeric)

we removed duplicate diagnosis codes and converted the remaining 3,070 diagnosis codes into digital numbers from 1 to 3,070. The **Supplementary Table 1** shows the diagnostic information of two patients.

Then, for the health condition of patients admitted to hospital, we used the minimum, maximum, median, mean, and variance value as the 5-tuple features of each laboratory indicator, and designed a symptom identification method based on text analysis of patient discharge reports, including rule setting, text segmentation, text extraction, abbreviation dictionary construction, negative word recognition, case unification, word segmentation, stop word removal, and external symptom dictionary embedding (**Supplementary Figure 1**). Additionally, we added related indicators to measure patients' severity, such as AIDS, hematologic malignancy, metastatic cancer SOFA, SAPS, and SAPS-II. Finally, we obtained 120 features of the health condition of sepsis patients in the experimental dataset, as shown in **Table 1**.

Method

Figure 2 shows the proposed UD identification and prediction method (UDIPM), which uses four types of information from EMRs. We adopt diagnostic information to identify the UD, and use demographic information, symptom information, and laboratory examination information to predict the UD. First, we apply a set of similarity measure methods to a large number of patients by embedding the semantic relation of the ICD classification system (Task 1 in **Figure 2**). Second, we apply a clustering algorithm to the similarity matrix to divide patients into different groups, and further obtain the exemplar and core patients of each cluster (Task 2 in **Figure 2**). Third, we extract the typical diagnosis code co-occurrence patterns (TDCCoP)

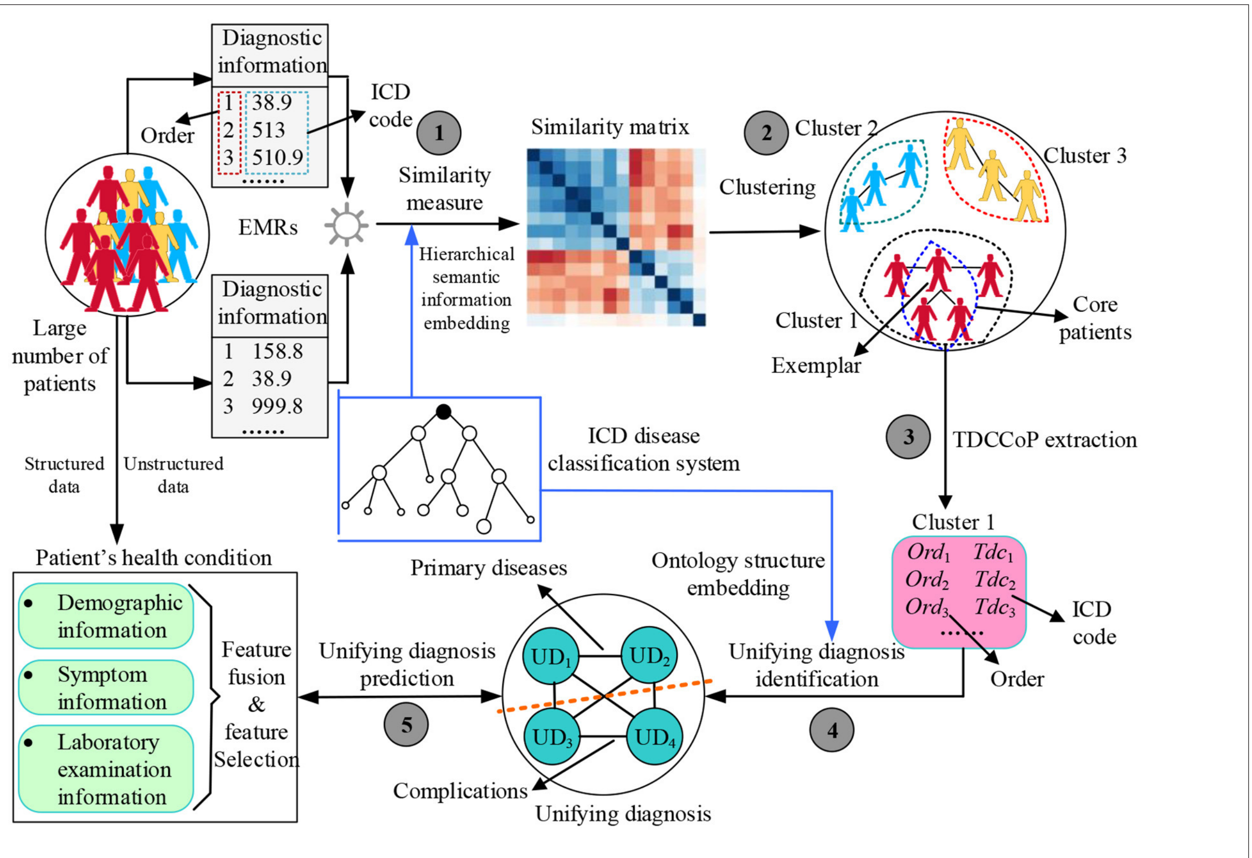

FIGURE 2 | Research framework for applying the proposed UDIPM to EMRs.

from each cluster by defining a threshold and a sorting function (Task 3 in **Figure 2**). Fourth, we combine the visual analysis and conditional co-occurrence matrix (CCoM) to identify the UD by selecting the optimal segmentation (Task 4 in **Figure 2**). Finally, after obtaining the health condition of the patient admitted to hospital, we obtain a UD prediction using multi-class classification methods (Task 5 in **Figure 2**).

Patient Similarity Measure Method

Many methods exist for measuring patient similarity (29, 30). In this study, considering the semantic relations of diagnosis codes in the ICD ontology structure, we adopt a set similarity measure method. First, we define patient diagnostic information as a series of ordered diagnosis codes. Then we reconstruct the ontology structure based on a disease classification system to easily measure patient similarity. Finally, we describe the process of the set similarity method, including the information content (IC) measure of diagnosis codes, diagnosis code similarity measure, and diagnosis code set similarity measure.

Patient's Diagnostic Information Representation

Diagnostic information refers to a record of disease diagnosis made by clinicians based on the health condition of a patient admitted to hospital. It is stored in the patient's EMR data in the form of a diagnosis code (e.g., ICD-9 and ICD-10). Because of the prevalence of disease complications, a patient's EMR is typically annotated using multiple disease codes, and these codes have a certain priority (i.e., order). The higher the priority of the diagnosis code is, the more central and important the disease is for this patient, then the weaker conversely. Thus, patient diagnostic information can be represented as

$$D = \{(dc_1, Ord(dc_1)), (dc_2, Ord(dc_2)), \cdots, (dc_i, Ord(dc_i)), \cdots\}, \tag{1}$$

where dc_i and $Ord(dc_i)$ represent the i-th diagnosis code and its order, respectively.

Ontology Structure Construction

We automatically construct a five-level ICD-9 ontology structure, shown in **Figure 3**, in which level-0 is the virtual root node, level-1 has 19 chapters, level-2 has 129 sections, level-3 has ∼1,300 categories (**Supplementary Figure 2**), and the last two levels are expanded to 10 types of sub-nodes under each node. For example, level-4 contains 550.0, 550.1, 550.2 (virtual code), 550.3 (virtual code), … and 550.9, and level-5 includes 550.10, 550.11, 550.12, 550.13, 550.14 (virtual code), … 550.19 (virtual code). More importantly, the actual diagnosis codes of patients belong to the ICD-9 ontology structure, whereas the virtual codes are only used to construct a complete ICD ontology structure and do not play a role in the actual similarity measure.

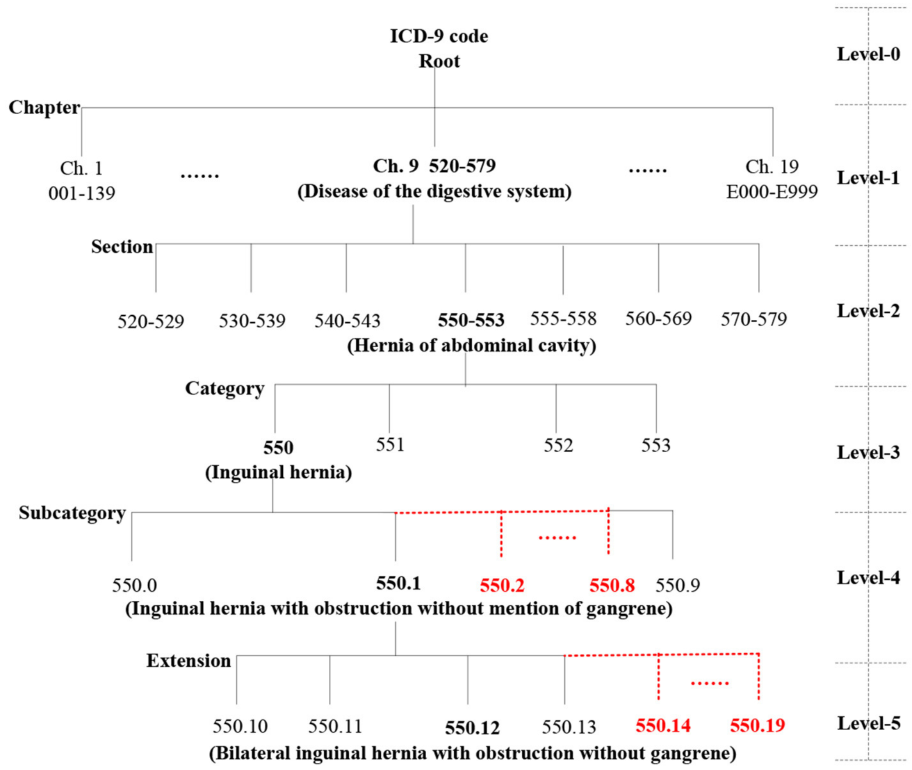

FIGURE 3 | Local ontology structure of ICD-9 codes.

Set Similarity Measure

Information Content Measure of Diagnosis Codes

In the ICD-9 ontology structure, each code represents a concept, and there is semantic similarity between classification concepts. Additionally, concepts on the same branch are more similar than those on different branches. Thus, we use the level depth measure method of the hierarchical tree (29), that is, we assign a value to each level of the ICD-9 ontology structure; the deeper the concept level, the larger the value. For an ICD-9 code dc_i, the IC is defined as

$$IC(dc_i) = level(dc_i \rightarrow Root), \tag{2}$$

where *Root* is the virtual root node and the function *level*(.) denotes the level depth from the ICD-9 code d_i to the root node. Intuitively, the IC of the root node (level-0) is 0, the ICs of a chapter (level-1), section (level-2), category (level-3), subcategory (level-4), and extension (level-5) are 1, 2, 3, 4, and 5, respectively.

Code-Level Similarity Measure

For the IC of codes, there are several approaches to measure code-level similarity. We use the least common ancestor (LCA) of two codes to measure the similarity of diagnosis codes, defined as

$$s(dc_i, dc_j) = \frac{2IC(LCA(dc_i, dc_j))}{IC(dc_i) + IC(dc_j)}, \tag{3}$$

where dc_i and dc_j are two diagnosis codes, and $LCA(dc_i, dc_j)$ is the LCA of dc_i and dc_j. If $dc_i = dc_j$, then $LCA(dc_i, dc_j) = dc_i = dc_j$, and $IC[LCA(dc_i, dc_j)] = IC\ (dc_i) = IC\ (dc_j)$. If $dc_i \neq dc_j$ and $LCA(dc_i, dc_j) = Root$, then $IC[LCA(dc_i, dc_j)] = 0$.

To make this concept easier to understand, we provide a simple example in **Figure 4A**. Thus, $LCA(550.12, 550.13) = 550.1$, $LCA(541, 550.13) = 520\text{–}579$, $s = s_1(550.12, 550.13) = 2IC(550.1)/[IC(550.12) + IC(550.13)] = 2 * 4/(5 + 5) = 0.8$.

Code Set-Level Similarity Measure

In the EMR dataset, patient diagnostic information is typically a set of diagnosis codes. Thus, patient similarity can be transformed into the similarity of the diagnosis code set. Generally, for binary code-level similarity, we can use classical methods, such as Dice, Jaccard, cosine, and overlap, to calculate set-level similarity. However, these methods cannot fully embed semantic similarity. Thus, we use the most similar concept pair's average value to measure the set-level similarity (29), and the formula is defined as

$$S(D'_i, D'_j) = 1 - \frac{\left(\sum_{dc_{ig} \in D'_i} \min_{dc_{jh} \in D'_j}(1 - s(dc_{ig}, dc_{jh})) + \sum_{dc_{jh} \in D'_j} \min_{dc_{ig} \in D'_i}(1 - s(dc_{jh}, dc_{ig}))\right)}{|D'_i| + |D'_j|}, \tag{4}$$

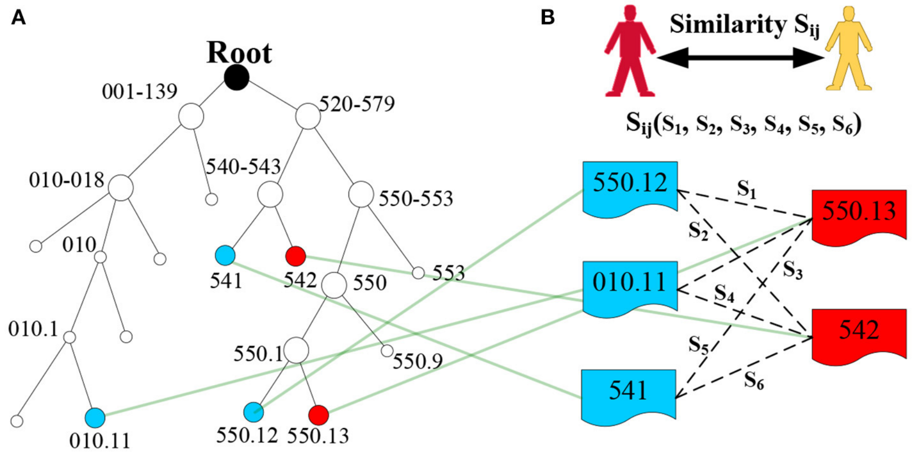

FIGURE 4 | Example of LCA generation in the ICD-9 ontology structure. **(A)** Denotes the ICD-9 ontology structure, and **(B)** denotes the diagnosis codes of two patients.

where $\boldsymbol{D}_i'$ and $\boldsymbol{D}_j'$ are the diagnostic information of patient i and patient j, respectively, which does not consider the order of diagnosis codes; that is, $\boldsymbol{D}_i'=\{dc_{i1}, dc_{i2},\ldots, dc_{ig},\ldots\}$ and $\boldsymbol{D}_j'=\{dc_{j1}, dc_{j2},\ldots, dc_{jh},\ldots\}$. $|\boldsymbol{D}_i'|$ and $|\boldsymbol{D}_j'|$ are the number of diagnosis codes for patient i and patient j, and dc_{ig} and dc_{jh} are the g-th diagnosis code of patient i and the h-th diagnosis code of patient j, respectively. Finally, we obtain the similarity $\mathbf{S}_{ij}$ of the two patients (**Figure 4B**), and similarity matrix $\mathbf{S}$ for all patients in the EMRs using a set similarity measure method. The pseudocode of the patient similarity measure method is presented in **Algorithm 1**.

Algorithm 1 | Patient similarity measure method.

```
Input: $\boldsymbol{D}_i'=\{dc_{i1}, dc_{i2},. ..,dc_{ig},...\}, i = 1, 2,\ldots, N$
Output: Similarity matrix $\mathbf{S}_{N*N}$
1. Construct the ICD ontology structure
2. For $i = 1: N$ do
    For $j = i + 1: N$ do
        Compute $IC(dc_{i1}, dc_{i2}, ...,dc_{ig},...)$, $IC(dc_{j1}, dc_{j2},..., dc_{jh},...)$, and
        diagnosis code similarity $s(dc_{ig}, dc_{jh}) = 2IC(LCA(dc_{ig}, dc_{jh}))/(IC(dc_{ig}) + IC(dc_{jh}))$
        based on the ICD ontology structure, compute set similarity $S(\boldsymbol{D}_i', \boldsymbol{D}_j')$ using Eq. 4
3. Obtain the similarity matrix $\mathbf{S}_{N*N}$ for $N$ patients
```

Patient Clustering Algorithm

A clustering algorithm aims to divide patients into multiple groups based on the similarity matrix **S**, requiring that patients in the same group are as similar as possible, and patients in different groups are as dissimilarity as possible (31, 32). In this study, considering the advantages, such as not predefining the number of clusters, the real existence of exemplars, and much lower error, we adopt affinity propagation (AP) clustering (33, 34).

AP clustering determines the number of clusters by controlling the input exemplar preferences (p), where p is more robust than K because p monotonically controls the perception granularity. Generally, p depends on the similarity matrix $\mathbf{S}_{N*N}$, number of input patients (N), and p coefficient (p_{coe}), which is represented as

$$p = median(\mathbf{S}) - p_{coe} * N. \tag{5}$$

After patients are clustered, we identify K clusters ($\boldsymbol{C}_1, \boldsymbol{C}_2,\ldots, \boldsymbol{C}_K$), and define the popularity (i.e., support) of each cluster as

$$Support(C_k) = \frac{\sum_{j\in\{1,2,\cdots,N\}} \lambda(C(D'_j), E(C_k))}{N}, k=1, 2, \cdots, K, \tag{6}$$

where $\boldsymbol{C}(D_j')$ represents the cluster to which patient j belongs and $E(\boldsymbol{C}_k)$ denotes the exemplar of $\boldsymbol{C}_k$. $\lambda(.)$ is an indicator function; if patient j belongs to $\boldsymbol{C}_k$, then $\lambda[\boldsymbol{C}(D_j'), E(\boldsymbol{C}_k)] = 1$; otherwise, $\lambda[\boldsymbol{C}(D_j'), E(\boldsymbol{C}_k)] = 0$.

Additionally, we obtain the sum of similarities (SS), which is an important indicator used to evaluate clustering performance. The SS depends on the similarity matrix $\mathbf{S}_{N*N}$, number of input patients (N), number of clusters (K), and corresponding exemplars, which is represented as

$$SS(K) = \sum_{i=1}^{K} \sum_{D'_j\in C_i} S(D'_j, E(C_i)). \tag{7}$$

Generally, the larger the SS value, the better the clustering performance. The pseudocode of the patient clustering algorithm is presented in **Algorithm 2**.

TDCCoP Extraction Method

In our previous studies, we proved that defining the core zone of a cluster is an effective approach to extract stable clustering results (35). Additionally, considering the complex semantic relations

Algorithm 2 | Patient clustering algorithm.

```
Input: $\mathbf{S}_{N*N}$, $p_{coe}$, step size $\boldsymbol{\varepsilon}$
Output: Optimal clustering number $K^{*}$, $E(\boldsymbol{C}_k)$, support $(\boldsymbol{C}_k)$, SS($K^{*}$)
1. Initialize $\mu = 1$, $p_{coe}(\mu) = p_{coe}$=0, $\boldsymbol{\varepsilon}$
2. Run the AP clustering algorithm with $\mathbf{S}_{N*N}$ and $p$ ($p$ = median($\mathbf{S}$)-$p_{coe}(\mu)$*$N$)
3. Return the clustering number $K(\mu)$
4. While $K(\mu) < N$ and $K(\mu) > 1$ do
     $\mu = \mu+1$, $p_{coe}(\mu) = p_{coe}(\mu - 1) + \boldsymbol{\varepsilon}$
     $p$ = median($\mathbf{S}$) – $p_{coe}(\mu)$ * $N$
     Run the AP clustering algorithm with $\mathbf{S}_{N*N}$ and $p$
     Return the clustering number $K(\mu)$ and $p_{coe}(\mu)$
5. Compute the distance $dp_{coe}(K) = max\,[p_{coe}(\mu_i)] – min\,[p_{coe}(\mu_j)]$ for
   the same $K$
6. Return the maximum $dp_{coe}(K)$ and the optimal clustering number $K^{*}$
7. Set $p_{coe}$=0.5 * {$max\,[p_{coe}(\mu_i)] + min\,[p_{coe}(\mu_j)]$} for $K^{*}$
8. Run the AP clustering algorithm with $\mathbf{S}_{N*N}$ and $p$ ($p$ = median($\mathbf{S}$) – $p_{coe}$* $N$)
9. Return $E(\boldsymbol{C}_k)$, support $(\boldsymbol{C}_k)$ using Eq. 6, and SS($K^{*}$) using Eq. 7
```

among different diagnosis codes, the feature of a cluster cannot be fully described when the diagnostic information (cluster center or exemplar) of only one patient is used. Thus, we also define the core zone of each cluster to select a group of patients (i.e., core patients) using the k-nearest neighbor method, and further extract typical diagnosis codes (TDCs). For cluster $\boldsymbol{C}_k$, the core zone is defined as

$$Core_k = \left\{ D'_j | S(D'_j, E(C_k)) \geq \tau \right\}, \tag{8}$$

where $E(\boldsymbol{C}_k)$ is the exemplar of cluster $\boldsymbol{C}_k$ and τ is a similarity threshold defined in advance, which aims to determine the number of core patients.

Then, for cluster $\boldsymbol{C}_k$, the occurrence probability of the diagnosis code dc_h can be represented as

$$Prob_k(dc_h) = \frac{\sum_{D'_j \in Core_k} \lambda(dc_h, D'_j)}{|Core_k|}, h = 1, \cdots, H, \tag{9}$$

where $|\text{Core}_k|$ denotes the number of core patients in cluster $\boldsymbol{C}_k$. $\lambda(.)$ is an indicator function; if the diagnostic information $\boldsymbol{D}'_j$ of patient j contains diagnosis code dc_h, then λ $(dc_h, \boldsymbol{D}'_j) = 1$; otherwise, λ $(dc_h, \boldsymbol{D}'_j) = 0$. H is the number of all diagnosis codes after duplicates are deleted.

After we calculate the probability of all diagnosis codes in the cluster $\boldsymbol{C}_k$, we define the TDC as

$$Tdc_h = \left\{ dc_h | Prob_k(dc_h) > \delta_1 \right\}, \tag{10}$$

where δ_1 is a threshold defined in advance to differentiate high-frequency and low-frequency diagnosis codes.

Based on all TDCs of the cluster $\boldsymbol{C}_k$, we further analyze the priority of TDCs by embedding the order of the patient diagnostic information, that is, for patient j, D_j= {[dc_{j1}, $Ord(dc_{j1})$], [dc_{j2}, $Ord(dc_{j2})$], [dc_{jh}, $Ord(dc_{jh})$], …} and $D_j{}' =$ {dc_{j1}, dc_{j2}, dc_{jh}, …}. Thus, the average order (AOrd) of TDC $Tdc_h$is defined as

$$AOrd(Tdc_h) = \frac{\sum_{D'_j \in Core_k, Tdc_h \in D'_j} Ord_{D_j}(Tdc_h)\lambda(Tdc_h, D'_j)}{\sum_{D'_j \in Core_k, Tdc_h \in D'_j} \lambda(Tdc_h, D'_j)},$$
$$h = 1, \cdots, H', \tag{11}$$

where H' is the number of TDCs in cluster $\boldsymbol{C}_k$ and $Ord_{Dj}(Tdc_h)$ denotes the order of TDC Tdc_h in the diagnostic information D_j of patient j. Generally, the smaller the AOrds of typical diagnostic codes, the more likely they are to be primary diseases.

Finally, after obtaining TDCs and their AOrds, we define a sorting function to determine TDCCoP, which is represented as

$$\begin{aligned} TDCCoP_k &= Sort((Tdc_1, AOrd_k(Tdc_1)), \cdots, (Tdc_{H'}, AOrd_k(Tdc_{H'}))) \\ &= \{(Tdc_1, Ord'(Tdc_1)), \cdots, (Tdc_{H'}, Ord'(Tdc_{H'}))\}, \end{aligned} \tag{12}$$

where Ord' (Tdc_h) is the new order of Tdc_h. For example, if cluster $\boldsymbol{C}_k$ has only three TDCs (e.g., Tdc_1, Tdc_2, and Tdc_3) and its AOrds are 5.3, 7.8, and 3.8, respectively, then after sorting, the $TDCCoP_k$ is {(Tdc_3, 1), (Tdc_1, 2), (Tdc_2, 3)}. The pseudocode of the TDCCoP extraction method is presented in **Algorithm 3**.

Algorithm 3 | TDCCoP extraction method.

```
Input: $C_k$, $E(\boldsymbol{C}_k)$, $Core_k$, $\boldsymbol{D}_j$, $\boldsymbol{D}'_j$, $H$, $k = 1, 2 \ldots, K$
Output: $TDCCoP_k$, $k = 1, 2, \ldots, K$
1. Initialize $k = 1$, $\tau$, $h = 1$, $\delta_1$, $i = 1$
2. For $k = 1$: $K$ do
     $Core_k$= {$\boldsymbol{D}'_j$|S($\boldsymbol{D}'_j$, $E(\boldsymbol{C}_k)$) $\geq \tau$}
     For $h = 1$:$H$ do
         $Prob_k(dc_h) = \sum_{Dj' \epsilon Corek} \lambda(dc_h, \boldsymbol{D}'_j)/|Core_k|$
         If $Prob_k(dc_h) > \delta_1$ then
             $Tdc_h \leftarrow dc_h$
         $i = i + 1$
     $H'$ $i$
     For $h = 1$:$H'$ do
         $AOrd_k$($Tdc_h$) = $\sum_{Tdch \epsilon Dj'} Ord(Tdc_h)/|\boldsymbol{D}'_j|$
     $TDCCoP_k$= $Sort$ (($Tdc_1$, $AOrd_k(Tdc_1)$),…, ($Tdc'_H$, $AOrd_k(Tdc'_H)$))
3. Return $TDCCoP_k$, $k = 1, 2, \ldots, K$
```

UD Identification Method

To identify a UD, categorizing the TDCCoP of each cluster reasonably according to the disease taxonomy is a critical step. In this study, we propose a UD identification method, as shown in **Figure 5**. Specifically, for the $TDCCoP_k$of cluster k, we first visualize all TDCs in the reconstructed ICD ontology structure, and mark their orders. Then we use the LCA method to categorize these codes, and define their LCA and the corresponding orders. Furthermore, we calculate the CCoM

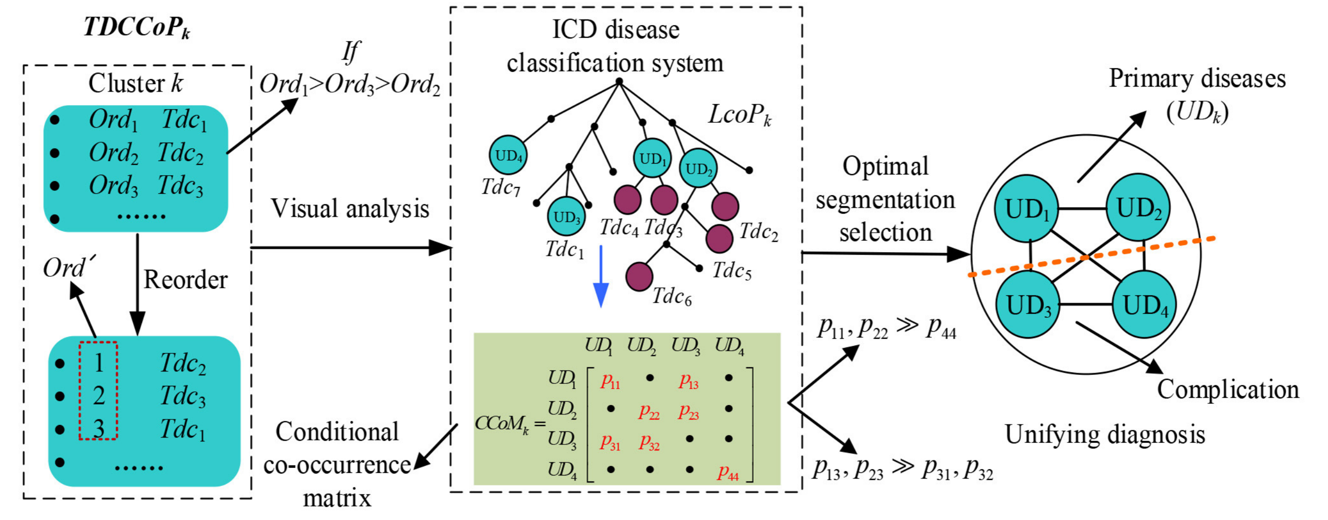

FIGURE 5 | Proposed UD identification method.

using patient diagnostic information to select the optimal segmentation between primary diseases and complications. Finally, we regard the identified primary diseases as the UD.

First, we define the LCA co-occurrence pattern (LCoP) of the $TDCCoP_k$ using visual analysis of the ICD ontology structure as

$$LCoP_k = \left\{d_i | d_i = LCA_{\{Tdc_1, Tdc_2, \cdots\} \in TDCCoP_k}(Tdc_1, Tdc_2, \cdots), d_i \neq Root\right\}. \quad (13)$$

Then we calculate the order of each $d_i$in $LCoP_k$ as

$$Ord(d_i) = min_{d_i = LCA(Tdc_1, Tdc_2, \cdots, Tdc_m)} (Ord'(Tdc_1), Ord'(Tdc_2), \cdots, Ord'(Tdc_m)), \quad (14)$$

where m is the number of TDCs in $LCoP_k$ whose LCA is d_i.

Additionally, considering the causal relation between d_i and $d_j$in $LCoP_k$, we define the conditional co-occurrence probabilities $p_k(d_j/d_i)$ and $p_k(d_i/d_j)$ as

$$\begin{aligned} p_k(d_j/d_i) &= Freq_k(d_j, d_i)/Freq_k(d_i) \\ p_k(d_i/d_j) &= Freq_k(d_i, d_j)/Freq_k(d_j) \end{aligned}, \quad (15)$$

where $Freq_k$ (d_i, d_j) and $Freq_k$ (d_j, d_i) denote the number of co-occurrences of d_i and d_j, respectively, and $Freq_k(d_i)$ denotes the number of occurrences of $d_i$in the cluster $\mathbf{C}_k$.

Thus, for all diagnosis codes in $LCoP_k$, we generate a CCoM $\mathbf{CCoM}_k$, where $\mathbf{CCoM}_k$ $(i, j) = p_k(d_j/d_i)$, $\mathbf{CoM}_k$ $(j, i) = p_k(d_i/d_j)$, and the diagonal entry $\mathbf{CCoM}_k$ $(i, i) = p_k(d_i) = Freq_k(d_i)/|\text{Core}_k|$. If $\mathbf{CCoM}_k$ $(i, j) >> \mathbf{CCoM}_k$ (j, i) or $\mathbf{CCoM}_k$ $(i, i) >> \mathbf{CCoM}_k$ (j, j) exist, then d_j is more prone to occur after the occurrence of d_i; thus, $d_i$is more likely to be a primary disease, whereas d_j will become a complication, and vice versa.

After analyzing the precedence relation of all diagnosis codes in $LCoP_k$ using $\mathbf{CCoM}_k$, we obtain the optimal segmentation between primary diseases and complications, and define the UD of cluster k as

$$UD_k = \left\{d_i | d_i \in LCoP_k, d_i \neq Complication\right\}, \quad (16)$$

where $\boldsymbol{UD}_k$ is a set of primary diseases. The pseudocode of the UD identification method is presented in **Algorithm 4**.

Algorithm 4 | UD identification method.

```
Input: $TDCCoP_k$, $Core_k$, $\mathbf{D}_j$, $\mathbf{D}_j'$, $k = 1, 2, \ldots, K$, ICD ontology structure
Output: $\boldsymbol{UD}_k$, $k = 1, 2, \ldots, K$
1. Initialize $i = 1$, call the ICD ontology structure
2. For $k = 1: K$ do
     While $Tdc\varepsilon\ TDCCoP_k$ do
         $d_i = LCA\ (Tdc_1, Tdc_2, \ldots)$
         $Ord(d_i) = min\ (Ord'(Tdc_1), Ord'(Tdc_2), \ldots)$
         $i = i + 1$
     $I \leftarrow i$
     For $i_1 = 1{:}I$ do
         For $i_2 = i_1 + 1{:}I$ do
             $p_k(d_{i1}) = \sum_{di1=LCA(Tdc1,\ldots Tdcg)} \lambda(Tdc_g, \mathbf{D}_j')/|Core_k|$
             $p_k(d_{i2}) = \sum_{di2=LCA(Tdc1,\ldots Tdch)} \lambda(Tdc_h, \mathbf{D}_j')/|Core_k|$
             $p_k(d_{i2}/d_{i1}) = (\sum_{di1=LCA(\ldots Tdcg),di2=LCA(\ldots Tdch)} \lambda(Tdc_g, Tdc_h, \mathbf{D}_j')/|Core_k|)/p_k(d_{i1})$
             $p_k(d_{i1}/d_{i2}) = (\sum_{di1=LCA(\ldots Tdcg),di2=LCA(\ldots Tdch)} \lambda(Tdc_g, Tdc_h, \mathbf{D}_j')/|Core_k|)/p_k(d_{i2})$
         If $p_k(d_{i1}) >> p_k(d_{i2})$ || $p_k(d_{i2}/d_{i1}) >> p_k(d_{i1}/d_{i2})$ || $Ord(d_{i1}) << Ord(d_{i2})$ then
             $\boldsymbol{UD}_k \leftarrow d_{i1}$
         Else
             $\boldsymbol{UD}_k \leftarrow d_{i2}$
3. Return $\boldsymbol{UD}_k$, $k = 1, 2, \ldots, K$
```

UD Prediction Method

After identifying the UD, we further study the prediction task based on the health condition of a patient admitted to hospital,

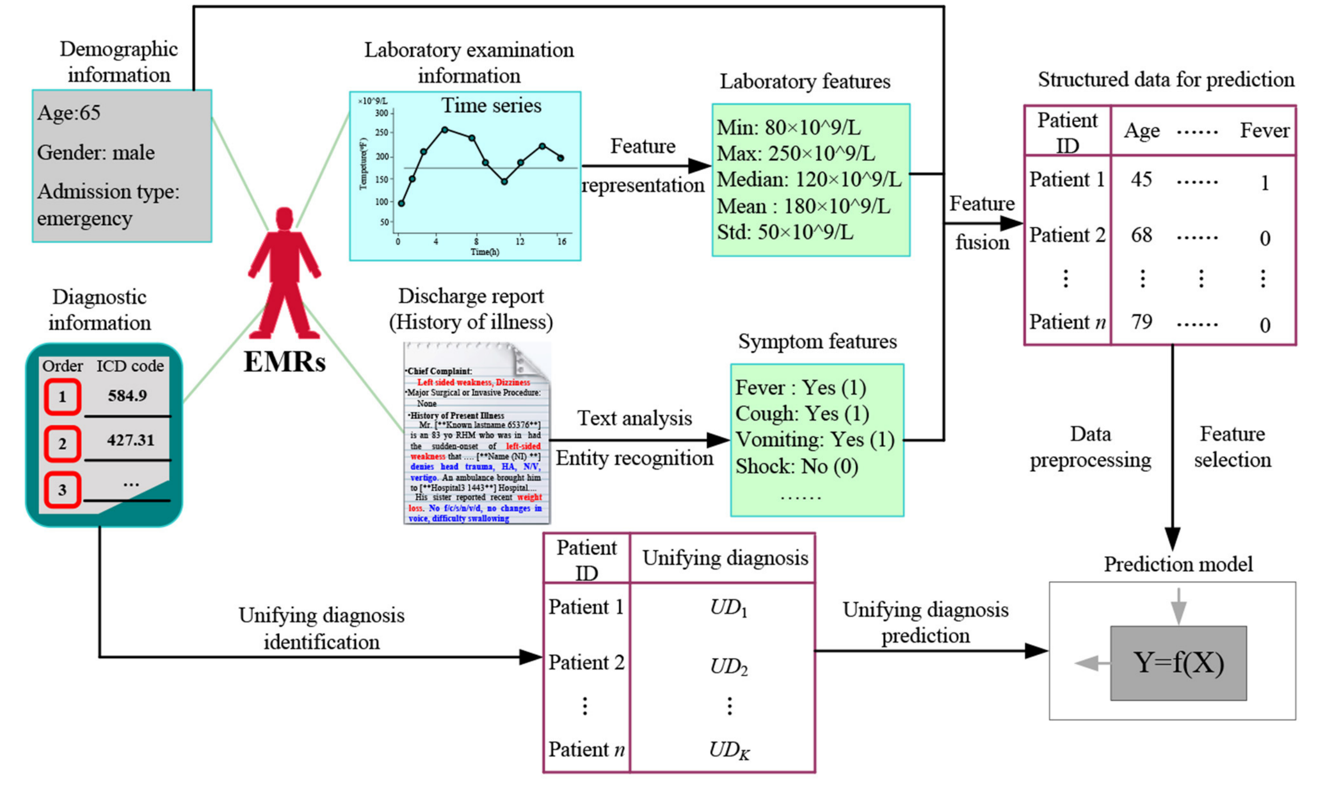

FIGURE 6 | Proposed UD prediction method.

exploring the important features to assign the most possible UDs to new patients. **Figure 6** shows the proposed UD prediction method. First, we extract three categories of features using time series feature representation and text analysis methods, and fuse them in structured data for further prediction. Then after data pre-processing and feature selection, we label all patients with a UD. Finally, we adopt classical prediction models to perform the UD prediction task.

Patient's Health Condition Representation

The health condition of a patient admitted to hospital includes demographic information, symptom information, and laboratory examination information, which play crucial roles for clinicians in diagnosing disease types, evaluating disease severity, and designing a treatment regimen.

Demographic Information

Demographic information mainly includes the date of birth, age, gender, admission type, marital status, occupation, and residence, defined as

$$De = \left\{De^{Age}, De^{Gender}, De^{Admission\ Type}, De^{Marital\ Status}, \cdots\right\}. \tag{17}$$

Symptom Information

Symptom information is recorded in the chief complaint and history of present illness in the form of text, where the chief complaint is the most painful part of the disease process, including the main symptoms and onset time. The history of present illness describes the entire process for the patient after suffering from diseases, including occurrence, development, evolution, diagnosis, and treatment. Thus, the patient's symptom information can be represented as.

$$Sy = \left\{Sy^{Fever}, Sy^{Weakness}, Sy^{Diarrhea}, \cdots\right\}. \tag{18}$$

Laboratory Examination Information

Laboratory examination refers to an indirect judgment of the health condition as a result of measuring specific components of blood and body fluids using instruments. Laboratory indicators typically have the characteristics of a time series, particularly for patients in the ICU. Thus, we use the minimum value, maximum value, median value, mean value, and variance of laboratory indicators to represent the time series, defined as

$$\begin{aligned} LE = \{\{(&min(LE^{WBC}), max(LE^{WBC}), med(LE^{WBC}), \\ &mean(LE^{WBC}), var(LE^{WBC})\}, \cdots\} \end{aligned} \tag{19}$$

Finally, we obtain the health condition of a patient admitted to hospital using a feature fusion method, that is, $\boldsymbol{X} = \{\boldsymbol{De}; \boldsymbol{Sy}; \boldsymbol{LE}\}$.

Information Gain-Based Feature Selection

Before predicting the UD, to remove noisy data, reduce the complexity and dimensionality of the dataset, and achieve accurate results, it is essential to apply feature selection methods to identify useful features. Therefore, feature selection is an important step that improves the clarity of the data and decreases the training time of prediction models (4). In this study, we use the information gain (IG) method to measure the importance of features and eliminate some irrelevant features. Then we compute the IG of feature x_i as

$$\begin{aligned} IG(x_i) &= H(Y) - H(Y/x_i) \\ &= -\sum_{k=1}^{K} P(y_k)\log P(y_k) + \sum_{k=1}^{K} P(y_k/x_i)\log P(y_k/x_i), \end{aligned} \tag{20}$$

where feature $x_i \epsilon \boldsymbol{X}$, $\boldsymbol{Y} = \{\boldsymbol{UD}_1,\ldots, \boldsymbol{UD}_k,\ldots, \boldsymbol{UD}_K\}$, $y_k \epsilon \boldsymbol{Y}$, $H(\boldsymbol{Y})$, and $H(\boldsymbol{Y}/x_i)$ denote the information entropy and conditional information entropy given feature x_i for a UD classification, and $P(y_k)$ and $P(y_k/x)$ denote the probability of y_k and condition probability of y_k given feature x_i, respectively.

Thus, we obtain the important features as

$$X' = \left\{x_i | IG(x_i) > \delta_2\right\}, \tag{21}$$

where δ_2 is a threshold defined in advance to differentiate the important and unimportant features using the IG method.

Prediction Model Establishment

After obtaining the feature representation and UD result of each patient, we generate a standard dataset (Y and X') and establish a prediction model [$Y = f\ (X')$]. In this study, we apply five classifiers to achieve a UD prediction: logistic regression, decision tree, random forest, SVM, and extreme gradient boosting (XGBoost). In the prediction process, we adopt the Z-fold cross-validation (CV) method, which randomly partitions the initial dataset into Z mutually exclusive subsets, and perform training and testing Z times. We set Z to 5 or 10. Then we compute the average CV error to determine the prediction model as

$$CVError_Z = \frac{1}{Z}\sum_{z=1}^{Z} L_z = \frac{1}{Z}\sum_{z=1}^{Z}\frac{1}{m_z}\sum_{j=1}^{m_z}(\hat{y}_j - y_j)^2, \tag{22}$$

where L_z and m_z are the average CV error and number of the z-th testing dataset, and y_j and $\hat{y}_j$ are the real and predicted UDs of the j-th patient, respectively.

Additionally, we identify distinctive features of different unifying diagnoses by analyzing the feature importance ranking results.

Parameter Setting

In our experiment, we set 5 parameters in advance. First, we set p_{coe}in Eq. 5 to select the number of clusters, and then τ in Eq. 8, which is a similarity threshold to determine the number of core patients (i.e., |Core|). We discuss both parameters based on the stability of the experimental results. We set δ_1 in Eq. 10 to 0.3 to obtain TDCs, and δ_2 in Eq. 21 to 0.005 to select the important features. We set the last parameter Z in Eq. 22 to 10 to perform the 10-fold CV method. In particular, before UD prediction, we used data pre-processing methods, that is, data normalization and smoothing for imbalanced classes.

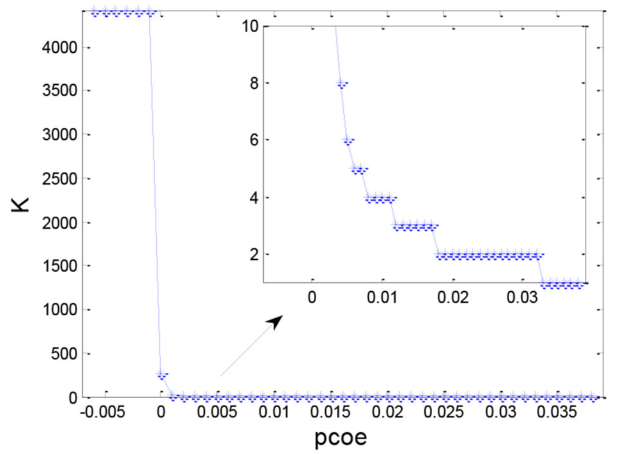

FIGURE 7 | Distribution of the number of clusters for different values of p_c.

RESULTS

Selection of the Cluster Number

After obtaining the set similarity measure based on the ontology structure for 4,418 sepsis patients, we obtained the similarity matrix **S** and used the AP clustering algorithm to divide all the patients into multiple groups. **Figure 7** shows the distribution of the number of clusters under different values of p_c. Generally, the number of clusters decreased as the preference coefficient increased. The most stable number of clusters was two when p_c ranged from 0.018 to 0.032. Thus, we selected two clusters (p_c = 0.025) to identify TDCs and extract TDCCoPs from each cluster.

Stability Analysis of TDCs

After applying the AP clustering algorithm, we first divided the 4,418 sepsis patients into two clusters, where cluster 1 and 2 contained 1,391 and 3,027 patients with a support of 31.48% and 68.52%, respectively. Then we analyzed the stability of the TDCs in Eq. 10 using a set of different numbers of core patients in Eq. 8 (|Core|=100, 200, 400, 500, 800, and all patients), as shown in **Figure 8**, **Supplementary Figure 3**.

From the distribution of TDCs in **Figure 8**, **Supplementary Figure 3**, the results showed that the stable range of core patients was from 400 to 800 (five codes in cluster 1 and 12 codes in cluster 2) because the number of TDCs and their distributions were approximately coincident. Specifically, compared with the stable TDCs, more TDCs were identified when the number of core patients was set to 100 and 200 (14 codes in cluster 2), such as the digital number 71 (276, disorders of fluid electrolyte and acid-base balance) and digital number 490 [V58.610, long-term (current) use of anticoagulants] (**Supplementary Figures 3A,B**). Digital number 99 (995.91, sepsis) was identified in cluster 1, and another three codes (486, 276.2, and 250) were not identified in cluster 2

(**Supplementary Figure 3E**) when we used all patients in the two clusters to extract TDCs. Thus, in the next experiment, we set the number of core patients to 800 to extract the TDCCoPs.

TDCCoP Extraction From Each Cluster

Using the clustering results, we finally determined two clusters, selected 800 core patients from each cluster, and set δ to 0.3 in Eq. 10 to identify TDCs and extract TDCCoPs. **Figure 9** shows the co-occurrence relation and AOrd of all TDCs in two TDCCoPs, and **Table 2** provides a detailed description of all TDCs in the two TDCCoPs.

To summarize, the experimental results indicated that there were 12 types of TDCs in the two TDCCoPs, where TDCCoP_1 and TDCCoP_2 had 5 and 12 codes, respectively. Specifically, the two TDCCoPs had similarities and differences. There were three similarities: (1) Five types of TDCs were the same, that is, 518.81, 38.9, 785.52, 584.9, and 995.92. (2) The AOrds of all TDCs in the same TDCCoPs were similar, for example, the AOrds of four TDCs in TDCCoP_1 were all below 6, whereas those of the TDCs in TDCCoP_2 were over 7. (3) The TDCs 38.9 (septicemia), 785.52 (septic shock), and 995.92 (severe sepsis) had the highest occurrence probability in the two TDCCoPs. There were also three differences: (1) TDCCoP_2 identified more TDCs than TDCCoP_1. (2) The occurrence probabilities of TDCs in TDCCoP_1 were larger than those in TDCCoP_2. (3) The AOrds of the same TDC were different in the two TDCCoPs, for example, 518.81 (acute respiratory failure) in the two TDCCoPs was 4.145 and 7.665, respectively. Additionally, septicemia (38.9) was a high-frequency and primary disease in sepsis patients, which is a life-threatening complication that can occur when bacteria from another infection enters the blood and spreads throughout the body.

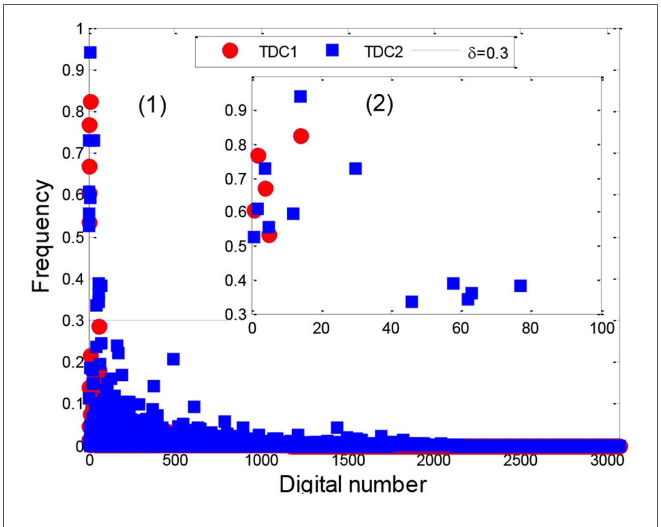

FIGURE 8 | Distribution of TDCs for 800 core patients.

Furthermore, using Eq. 12 and **Algorithm 3**, we extracted the TDCCOPs of the two clusters described in **Table 2**, that is, $\text{TDCCOP}_1 = \{(38.9, 1), (785.52, 2), (518.81, 3), (584.9, 4), (995.92, 5)\}$ and $\text{TDCCOP}_2 = \{(584.9, 1), (38.9, 2), (518.81, 3), (599.0, 4), (428.0, 5), (486.0, 6), (401.9, 7), (785.52, 8), (276.2, 9), (995.92, 10), (427.31, 11), (250.0, 12)\}$. Thus, from a reordering perspective, acute kidney failure, septicemia, and acute respiratory failure were probably the primary diseases in the two TDCCOPs.

UD Identification Based on TDCCOPs

After obtaining TDCCoPs, we visualized all the TDCs in the ICD-9 ontology structure. First, we categorized them using the LCA method to identify LCoPs using Eq. 13. Consider $TDCCoP_2$ as an example. The visualization result is shown in **Figure 10**. Clearly, we identified LCoP_2 with seven types of diseases, which are light green color, and computed the order of the new diseases using Eqs 13, 14: diseases of the genitourinary system (580–629, order: 1), septicemia (38.9, order: 2), diseases of the respiratory

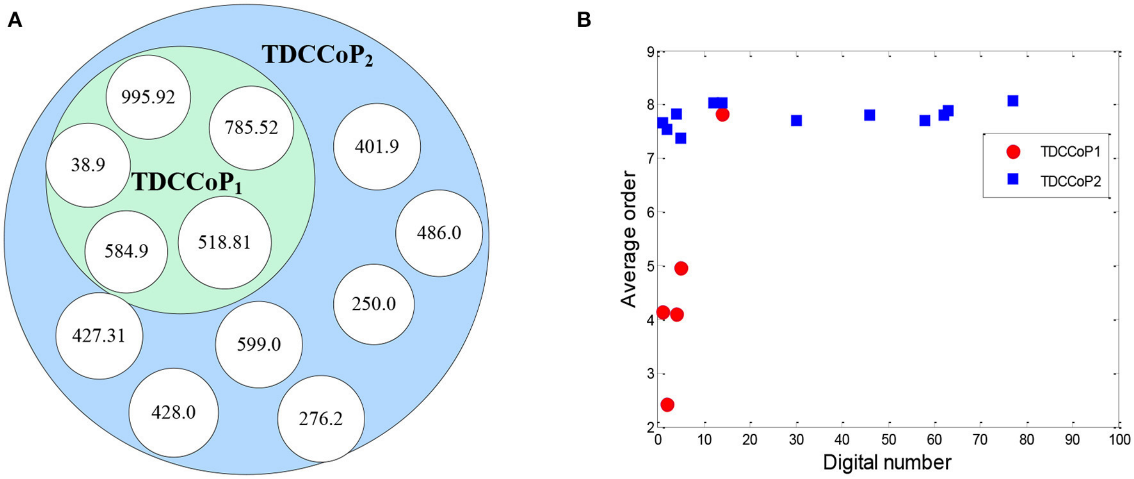

FIGURE 9 | Co-occurrence relation and AOrd of all TDCs. **(A)** Co-occurrence relation. **(B)** AOrd.

TABLE 2 | Detailed description of three *TDC*s.

TDCCOP	Digital number	TDC	Definition of diagnosis code	Occurrence frequency	Average order	Re-order
$TDCCOP_1$	1	518.81	Acute respiratory failure	0.604	4.145	3
(1391)	2	38.9	Septicemia	0.769	2.411	1
	4	785.52	Septic shock	0.669	4.090	2
	5	584.9	Acute kidney failure	0.534	4.956	4
	14	995.92	Severe sepsis	0.824	7.816	5
$TDCCOP_2$	1	518.81	Acute respiratory failure	0.526	7.665	3
(3027)	2	38.9	Septicemia	0.608	7.545	2
	4	785.52	Septic shock	0.729	7.813	8
	5	584.9	Acute kidney failure	0.554	7.377	1
	12	427.31	Atrial fibrillation	0.593	8.038	11
	14	995.92	Severe sepsis	0.941	8.031	10
	30	428.0	Congestive heart failure	0.729	7.703	5
	46	486.0	Pneumonia organism	0.334	7.805	6
	58	599.0	Urinary tract infection	0.389	7.701	4
	62	401.9	Essential hypertension	0.343	7.807	7
	63	276.2	Acidosis	0.360	7.875	9
	77	250.0	Diabetes mellitus without complication	0.383	8.062	12

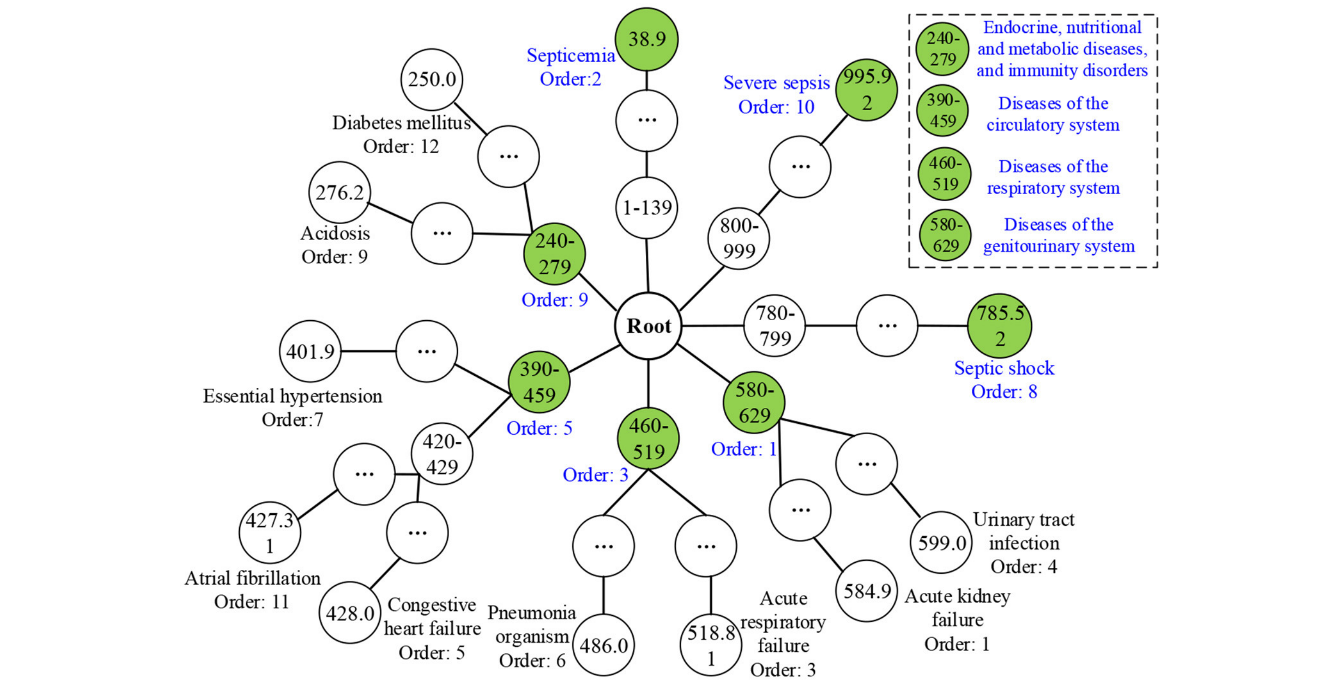

FIGURE 10 | $LCoP_2$ identified using the visualization of $TDCoP_3$ in the ontology structure.

system(460–519, order: 3), diseases of the circulatory system (390–459, order: 5), septic shock (785.52, order: 8), endocrine, nutritional, and metabolic diseases, and immunity disorders (240–279, order: 9), and severe sepsis (995.92, order: 10).

Then we calculated the $CCoM_2$ of the $LCoP_2$ based on the diagnostic information of 800 core patients in cluster 2, as described in **Table 3**. First, the conditional probabilities $p(\{390–459, 995.92\}/\{580–629, 38.9, 460–519\})$ colored red were significantly larger than the values $p(\{580–629, 38.9, 460–519\}/\{390–459, 995.92\})$ colored blue, which indicates that diseases of the genitourinary system (580–629, order: 1), septicemia (38.9, order: 2), and diseases of the respiratory system (460–519, order: 3) were more likely to be primary diseases, whereas diseases of the circulatory system (390–459, order: 5) and severe sepsis (995.92, order: 10) were probably complications.c Second, the orders of septic shock (785.52, order: 8) and endocrine, nutritional, and metabolic diseases, and immunity disorders (240–279, order: 9) were also larger than those of the first three diseases. Thus, diseases of the respiratory system (460–519, order: 3) and diseases of the circulatory system (390–459, order: 5) were likely to be the optimal segmentation between primary diseases and complications, and the first three diseases were considered to be the UD (UD_2) of cluster 2.

UD Prediction Based on Patient Admission Information

After we applied feature fusion and feature selection using the IG method, we further performed five classifications to predict a UD based on patient admission information and identify important features for the constructed prediction models. **Figure 11** shows the classification performance of the proposed UDIPM, including the area under the ROC curve (AUC), accuracy (Acc), precision (Pre), recall (Rec), and F1-score (F1), and **Figure 12** presents the 10 most important features identified using the random forest method (**Supplementary Figure 4**).

TABLE 3 | $CCoM_2$ of the $LCoP_2$.

$p_2(d_j/d_i)$	580-629	38.9	460-519	390-459	785.52	240-279	995.92
580-629 (1)	**0.75**	0.60	0.64	**0.92**	0.71	0.61	**0.94**
38.9 (2)	0.73	**0.61**	0.73	**0.93**	0.72	0.63	**0.94**
460-519 (3)	0.72	0.67	**0.66**	**0.92**	0.71	0.62	**0.94**
390-459 (5)	**0.74**	**0.61**	**0.66**	**0.93**	0.71	0.60	0.94
785.52 (8)	0.99	0.60	0.65	0.93	**0.73**	0.60	0.97
240-279 (9)	0.74	0.63	0.67	0.91	0.71	**0.61**	0.94
995.92 (10)	**0.74**	**0.61**	**0.66**	0.92	0.75	0.61	**0.94**

Values in brackets are the orders of the seven diseases, bold values on the master diagonal denote the occurrence probabilities of the seven diseases, and values in red and blue are conditional probabilities for distinguishing between primary diseases and complications.

The experimental results indicated that the proposed UDIPM achieved better prediction performance, where the AUC values were all above 0.8, except for the decision tree method. Similarly, the best Acc, Pre, Rec, and, F1 among all classifications was XGBoost, at ∼80%, followed by random forest, SVM, and logistic regression, whereas the decision tree was last, at ∼66%. Consider the random forest as an example. We obtained the feature importance results to better understand the prediction model. First, we found that demographic information (i.e., age) and laboratory examination information were more important than symptom information. Then some disease severity indicators were very important, such as SAPS and SAPS-II. Finally, the variance distribution (i.e., Var) of the laboratory examination indicators was more important than the mean, median, minimum, and maximum values. To summarize, the proposed UDIPM not only identified a UD from patient

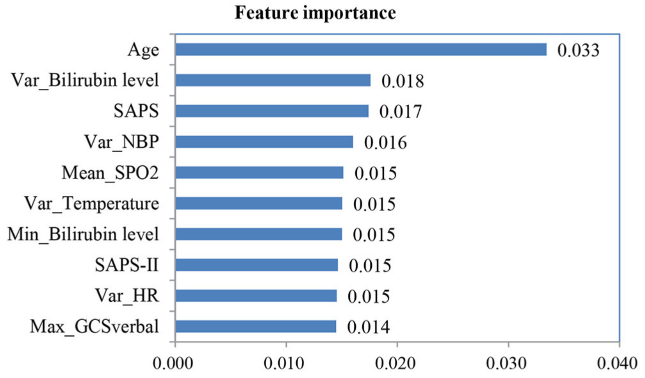

FIGURE 12 | Ten most important features using the random forest method.

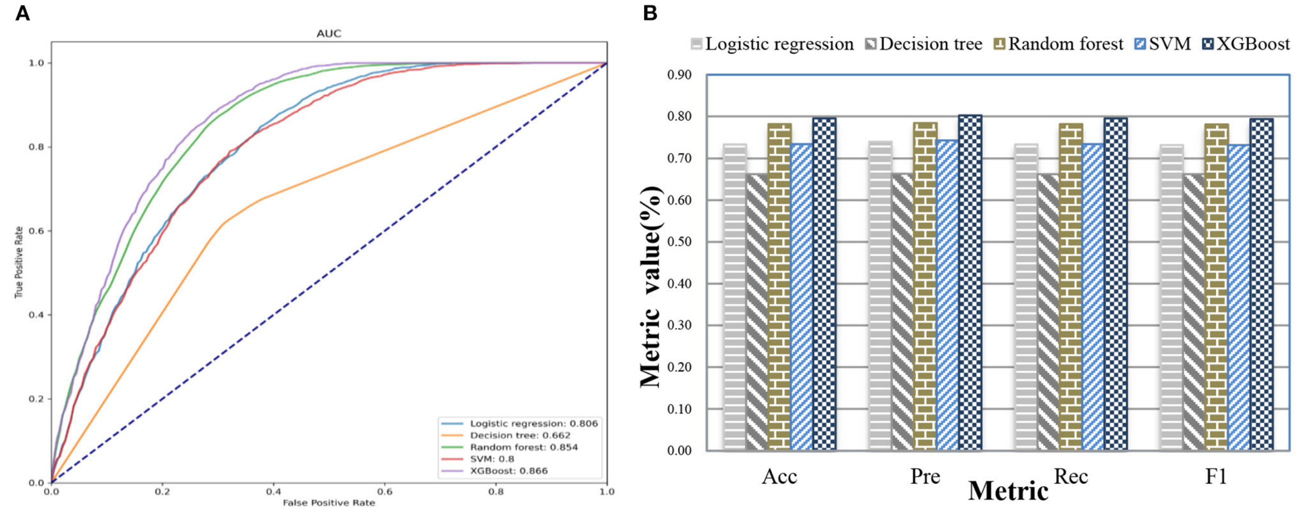

FIGURE 11 | Classification performance of the proposed UDIPM. **(A)** AUC. **(B)** Acc, Pre, Rec, and F1.

TABLE 4 | Evaluation methods and metrics used in our experiment.

Method name	Set similarity measure	Clustering	Classification
The proposed method (UDIPM)	Set similarity based on ontology	AP clustering	Logistic regression
Fusion method 1 (FM1)	Dice $= 2\left\|A\bigcap B\right\| / \|A\| + \|B\|$		Decision tree
Fusion method 2 (FM2)	Jaccrd $= \left\|A\bigcap B\right\| / \left\|A\bigcup B\right\|$		Random forest
Fusion method 3 (FM3)	Cosine $= \left\|A\bigcap B\right\| / \sqrt{\|A\| \cdot \|B\|}$		SVM
Fusion method 4 (FM4)	Overlap $= \left\|A\bigcap B\right\| / min\{\|A\|, \|B\|\}$		XGBoost

Metric	*SS* (Eq. 7)		AUC
			Acc = (TP + TN)/N
			Pre = TP/(TP + FP)
			Rec = TP/(TP + FN)
			F1 = 2Pre*Rec/(Pre + Rec)

A and B are the diagnosis code sets of two patients, the Dice method is the same as the proposed UDIPM when we do not consider the disease ontology structure and replace the code similarity with $s = (dc_i, dc_j) = \begin{cases} 1, & if\ dc_i = dc_j \\ 0 & otherwise \end{cases}$*, true positive (TP) and true negative (TN) measure the ability of classifier models to predict the UD, false positive (FP) and false negative (FN) identify the number of false predictions generated by the models, and we used FM to determine the prediction performance.*

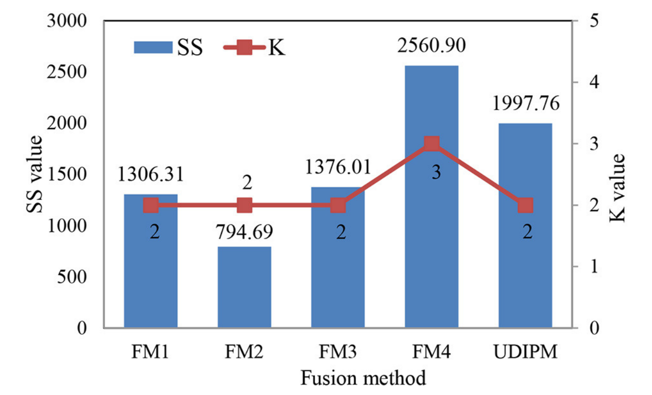

FIGURE 13 | Similarity measure and clustering results of different fusion methods.

diagnostic information but also predicted a UD based on the health condition of a patient admitted to hospital.

DISCUSSION

In this study, we conducted various experiments to demonstrate the efficiency of the proposed UDIPM when compared with other methods. Specifically, the proposed UDIPM fused three methods: a set similarity measure method, clustering, and classification algorithms. For the set similarity measure method, we selected Dice, Jaccard, cosine, and overlap as comparative methods, and used *SS* in Eq. 7 as a performance metric based on the AP clustering results. For the classification algorithms, we selected logistic regression, decision tree, random forest, SVM, and XGBoost. Additionally, we used AUC, Acc, Pre, Rec, and F1 as performance metrics to measure the effectiveness of the classification algorithms. The evaluation methods and metrics are described in detail in **Table 4**.

The detailed experimental results are shown in **Figure 13**, **Table 5**. Specifically, for the set similarity measure, we first selected the optimal number of clusters using AP clustering algorithms, and then computed the SS value based on the clustering results (**Algorithm 2**). The experimental results indicated that the optimal numbers of clusters for four FMs were 2, 2, 2, and 3 (**Supplementary Figure 5**), and the proposed UDIPM achieved the second-highest SS value of 1997.86; it was only below FM4 (**Figure 13**). The reason is that the SS value increased as the cluster number increased. Interestingly, although the similarities of FM1 and FM2 were different, they had the same clustering results.

TABLE 5 | Classification results of different fusion methods.

Fusion method	Classification algorithm	Metric				
		Acc	Pre	Rec	F1	AUC
FM1	Logistic regression	0.725	0.739	0.725	0.721	0.782
(Dice)	Decision tree	0.682	0.683	0.682	0.682	0.682
FM2	Random forest	0.779	0.782	0.779	0.778	0.851
(Jaccard)	SVM	0.722	0.763	0.722	0.711	0.778
	XGBoost	0.804	0.818	0.804	0.802	0.860
FM3	Logistic regression	0.734	0.743	0.734	0.732	0.804
(Cosine)	Decision tree	0.682	0.683	0.682	0.682	0.682
	Random forest	0.786	0.790	0.786	0.785	0.859
	SVM	0.736	0.752	0.736	0.732	0.801
	XGBoost	0.813	0.821	0.813	0.812	0.884
FM4	Logistic regression	0.465	0.437	0.421	0.411	0.628
(Overlap)	Decision tree	0.388	0.370	0.371	0.369	0.529
	Random forest	0.467	0.434	0.400	0.371	0.620
	SVM	0.471	0.384	0.404	0.350	0.626
	XGBoost	0.481	0.451	0.423	0.404	0.629
UDIPM	Logistic regression	0.733	0.740	0.733	0.732	0.806
	Decision tree	0.662	0.663	0.662	0.662	0.662
	Random forest	**0.782**	**0.784**	**0.782**	**0.781**	**0.854**
	SVM	0.734	0.743	0.734	0.732	0.800
	XGBoost	**0.795**	**0.803**	**0.795**	**0.794**	**0.866**

Bold values denote the first and second-highest performance using the UDIPM.

For the classification results obtained using the 10-fold CV method in **Table 5**, the proposed method achieved the second-highest performance using logistic regression, random forest, and SVM, and the third-highest performance using the decision tree and XGBoost. More importantly, all metrics of the proposed UDIPM were higher than those of FM4. Therefore, from the overall performance evaluation in combination with the set similarity measure, clustering, and classification, the UDIPM was an effective method for identifying and predicting a UD from EMRs.

Further, for all fusion methods, the results of performance comparison indicated that both XGBoost and random forest were superior to other classification algorithms in terms of the Acc, Pre, Rec, F1, and AUC. The main reason is that XGBoost and random forest are ensemble learning algorithms by combining multiple classifiers, which can often achieve more significant generalization performance than a single classifier. Specifically, XGBoost is an improved algorithm based on the gradient boosting decision tree, which can efficiently construct boosted trees and run in parallel. XGBoost works by combining a set of weaker machine learning algorithms to obtain an improved machine learning algorithm as a whole (36). XGBoost has been shown to perform exceptionally well in a variety of tasks in the areas of bioinformatics and medicine, such as the lysine glycation sites prediction for Homo sapiens (37), the chronic kidney disease diagnosis (38), and the risk prediction of incident diabetes (39). Also, random forest classifier is an ensemble algorithm, which combines multiple decorrelated decision tree prediction variables based on each subset of data samples (40). In general, random forest shows better performance in disease diagnosis than many single classifiers (41).

CONCLUSION

In this study, we proposed a UDIPM embedding the disease ontology structure to identify and predict a UD from EMRs to assist better coding integration of diagnosis in the ICU. We discussed many critical issues, including a formal representation of multi-type patient information, symptom feature extraction from an unstructured discharge report, ICD ontology structure reconstruction for semantic relation embedding, multi-level set similarity measure for generating a patient similarity matrix, number of cluster selections using AP clustering, stability of the extracted TDC and TDCCoP from each cluster, optimal split line determination for identifying a UD based on visual analysis and the CCoM of LCoP, feature fusion and selection using the IG-based method, and the performance evaluation of UD prediction using five classifiers. We verified the proposed UDIPM on 4,418 sepsis patients in the ICU extracted from the MIMIC-III database. The results showed that the highest stability cluster number and largest range of TDCs were 2 and 400–800, respectively, the UD of cluster 2 was diseases of the genitourinary system (580–629, order: 1), septicemia (38.9, order: 2), and diseases of the respiratory system (460–519, order: 3), and the best AUC and Acc, Pre, Rec, and F of the UD prediction were 0.866, 0.795, 0.803, 0.795, and 0.794, respectively, which were better than those of other fusion methods from the overall view of SS and prediction performance.

STUDY LIMITATIONS

The proposed UDIPM can identify and predict a UD from EMRs; however, there remain several topics for future work. First, the order of diagnosis codes should be considered in the patient similarity measure by way of different weights because of the importance of primary diseases. Then some state-of-the-art feature selection and classification models should be implemented to improve the prediction accuracies of the UD. Additionally, we hope to make progress on many of the valuable suggestions made by clinicians regarding our implemented method and experimental results.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: https://mimic.mit.edu/docs/iii/tables/.

AUTHOR CONTRIBUTIONS

JC, CG, and SD conceived and designed the study and revised the manuscript. JC and ML carried out the experiments and drafted the manuscript. All the authors read and approved the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant Nos. 71771034, 72101236, and 71421001), the Scientific and Technological Innovation Foundation of Dalian (Grant No. 2018J11CY009), the Henan Province Youth Talent Promotion Project (Grant No. 2021HYTP052), the Henan Province Medical Science and Technology Research Plan (Grant No. LHGJ20200279), and the Henan Province Key Scientific Research Projects of Universities (Grant No. 21A320035).

ACKNOWLEDGMENTS

We would like to thank the MIT Laboratory for Computational Physiology and collaborating research groups for providing the freely available database (MIMIC-III). We thank Maxine Garcia, Ph.D., from Liwen Bianji (Edanz) (www.liwenbianji.cn/) for editing the English text of a draft of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2021.793801/full#supplementary-material

REFERENCES

1. Herman J. The unifying diagnosis. *Scand J Prim Health.* (1994) 12:68–9. doi: 10.3109/02813439409003677
2. Xie J, Jiang J, Wang Y, Guan Y, Guo X. Learning an expandable EMR-based medical knowledge network to enhance clinical diagnosis. *Artif Intell Med.* (2020) 107:101927. doi: 10.1016/j.artmed.2020.101927
3. Sheikh A, Anderson M, Albala S, Casadei B, Franklin BD, Richards M, et al. Health information technology and digital innovation for national learning health and care systems. *Lancet Digit Health.* (2021) 3:e383–e96. doi: 10.1016/S2589-7500(21)00005-4
4. Ali F, El-Sappagh S, Islam SR, Kwak D, Ali A, Imran M, et al. A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. *Inf Fusion.* (2020) 63:208–22. doi: 10.1016/j.inffus.2020.06.008
5. Yadav P, Steinbach M, Kumar V, Simon G. Mining electronic health records (EHRs) A survey. *ACM Comput Surv.* (2018) 50:1–40. doi: 10.1145/3127881
6. Lin AL, Chen WC, Hong JC. Electronic health record data mining for artificial intelligence healthcare. *Artif Intell Med.* (2021) 133–50. doi: 10.1016/B978-0-12-821259-2.00008-9
7. Haque A, Milstein A, Fei-Fei L. Illuminating the dark spaces of healthcare with ambient intelligence. *Nature.* (2020) 585:193–202. doi: 10.1038/s41586-020-2669-y
8. Myszczynska MA, Ojamies PN, Lacoste AM, Neil D, Saffari A, Mead R, et al. Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nat Rev Neurol.* (2020) 16:440–56. doi: 10.1038/s41582-020-0377-8
9. Guo C, Chen J. Big data analytics in healthcare: data-driven methods for typical treatment pattern mining. *J Syst Sci Syst Eng.* (2019) 28:694–714. doi: 10.1007/s11518-019-5437-5
10. Piri S. Missing care: a framework to address the issue of frequent missing values; The case of a clinical decision support system for Parkinson's disease. *Decis Support Syst.* (2020) 136:113339. doi: 10.1016/j.dss.2020.113339
11. Wang S, Li X. Chang* X, Yao L, Sheng QZ, Long G. Learning multiple diagnosis codes for ICU patients with local disease correlation mining. *ACM T Knowl Discov D (TKDD).* (2017) 11:1–21. doi: 10.1145/3003729
12. Huang J, Osorio C, Sy LW. An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes. *Comput Meth Programs Biomed.* (2019) 177:141–53. doi: 10.1016/j.cmpb.2019.05.024
13. Gour N, Khanna P. Multi-class multi-label ophthalmological disease detection using transfer learning based convolutional neural network. *Biomed Signal Proces.* (2021) 66:102329. doi: 10.1016/j.bspc.2020.102329
14. Trigueros O, Blanco A, Lebena N, Casillas A, Perez A. Explainable ICD multi-label classification of EHRs in Spanish with convolutional attention. *Int J Med Inform.* (2021) 157:104615. doi: 10.1016/j.ijmedinf.2021.104615
15. Zhang Y, Yang Q. A survey on multi-task learning. *IEEE Trans Knowl Data Eng.* (2021). doi: 10.1109/TKDE.2021.3070203. [Epub ahead of print].
16. Yu K, Xie X. Predicting hospital readmission: a joint ensemble-learning model. *IEEE J Biomed Health.* (2019) 24:447–56. doi: 10.1109/JBHI.2019.2938995
17. Li T, Wang Z, Lu W, Zhang Q, Li D. Electronic health records based reinforcement learning for treatment optimizing. *Inf Syst.* (2022) 104:101878. doi: 10.1016/j.is.2021.101878
18. Chen PF, Wang SM, Liao WC, Kuo LC, Chen KC, Lin YC, et al. Automatic ICD-10 coding and training system: deep neural network based on supervised learning. *JMIR Med Inf.* (2021) 9:e23230. doi: 10.2196/23230
19. Sareen J, Olafson K, Kredentser MS, Bienvenu OJ, Blouw M, Bolton JM, et al. The 5-year incidence of mental disorders in a population-based ICU survivor cohort. *Crit Care Med.* (2020) 48:e675–e83. doi: 10.1097/CCM.0000000000004413
20. Johnson AE, Pollard TJ, Shen L, Li-Wei HL, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. *Sci Data.* (2016) 3:1–9. doi: 10.1038/sdata.2016.35
21. Diao X, Huo Y, Zhao S, Yuan J, Cui M, Wang Y, et al. Automated ICD coding for primary diagnosis via clinically interpretable machine learning. *Int J Med Inform.* (2021) 153:104543. doi: 10.1016/j.ijmedinf.2021.104543
22. Wu Y, Zeng M, Fei Z, Yu Y, Wu F-X, Li M. KAICD: a knowledge attention-based deep learning framework for automatic ICD coding. *Neurocomputing.* (2020) 469:376–83. doi: 10.1016/j.neucom.2020.05.115
23. Baumel T, Nassour-Kassis J, Cohen R, Elhadad M, Elhadad N. Multi-label classification of patient notes: case study on ICD code assignment. In: *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence.* New Orleans, LA (2018). p. 409–16.
24. Malhi GS, Bell E, Boyce P, Mulder R, Porter RJ. Unifying the diagnosis of mood disorders. *Aust N Z J Psychiatry.* (2020) 54:561–5. doi: 10.1177/0004867420926241
25. Sloan EA, Chiang J, Villanueva-Meyer JE, Alexandrescu S, Eschbacher JM, Wang W, et al. Intracranial mesenchymal tumor with FET-CREB fusion-A unifying diagnosis for the spectrum of intracranial myxoid mesenchymal tumors and angiomatoid fibrous histiocytoma-like neoplasms. *Brain Pathol.* (2021) 31:e12918. doi: 10.1111/bpa.12918
26. Liang JJ, Goodsell K, Grogan M, Ackerman MJ. LMNA-mediated arrhythmogenic right ventricular cardiomyopathy and charcot-marie-tooth type 2B1: a patient-discovered unifying diagnosis. *J Cardiovasc Electrophysiol.* (2016) 27:868–71. doi: 10.1111/jce.12984
27. Zhu Y, Zhang J, Wang G, Yao R, Ren C, Chen G, et al. Machine learning prediction models for mechanically ventilated patients: analyses of the MIMIC-III database. *Front Med.* (2021) 8:662340. doi: 10.3389/fmed.2021.662340
28. Kong G, Lin K, Hu Y. Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU. *BMC Med Inform Decis.* (2020) 20:1–10. doi: 10.1186/s12911-020-01271-2
29. Jia Z, Lu X, Duan H, Li H. Using the distance between sets of hierarchical taxonomic clinical concepts to measure patient similarity. *BMC Med Inform Decis.* (2019) 19:1–11. doi: 10.1186/s12911-019-0807-y
30. Jia Z, Zeng X, Duan H, Lu X, Li H, A. patient-similarity-based model for diagnostic prediction. *Int J Med Inform.* (2020) 135:104073. doi: 10.1016/j.ijmedinf.2019.104073
31. Park S, Xu H, Zhao H. Integrating multidimensional data for clustering analysis with applications to cancer patient data. *J Am Stat Assoc.* (2021) 116:14–26. doi: 10.1080/01621459.2020.1730853
32. Lopez-Martinez-Carrasco A, Juarez JM, Campos M, Canovas-Segura B. A methodology based on Trace-based clustering for patient phenotyping. *Knowl Based Syst.* (2021) 232:107469. doi: 10.1016/j.knosys.2021.107469
33. Chen J, Sun L, Guo C, Wei W, Xie Y, A. data-driven framework of typical treatment process extraction and evaluation. *J Biomed Inform.* (2018) 83:178–95. doi: 10.1016/j.jbi.2018.06.004
34. Liu Y, Liu J, Jin Y, Li F, Zheng T. An affinity propagation clustering based particle swarm optimizer for dynamic optimization. *Knowl Based Syst.* (2020) 195:105711. doi: 10.1016/j.knosys.2020.105711
35. Chen J, Sun L, Guo C, Xie Y. A fusion framework to extract typical treatment patterns from electronic medical records. *Artif Intell Med.* (2020) 103:101782. doi: 10.1016/j.artmed.2019.101782
36. Liu J, Wu J, Liu S, Li M, Hu K, Li K. Predicting mortality of patients with acute kidney injury in the ICU using XGBoost model. *PLoS ONE.* (2021) 16:e0246306. doi: 10.1371/journal.pone.0246306
37. Yu B, Qiu W, Chen C, Ma A, Jiang J, Zhou H, et al. SubMito-XGBoost: predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting. *Bioinformatics.* (2020) 36:1074–81. doi: 10.1093/bioinformatics/btz734
38. Ogunleye A, Wang Q-G. XGBoost model for chronic kidney disease diagnosis. *IEEE ACM T COMPUT BI.* (2019) 17:2131–40. doi: 10.1109/TCBB.2019.2911071
39. Wu Y, Hu H, Cai J, Chen R, Zuo X, Cheng H, et al. Machine learning for predicting the 3-year risk of incident diabetes in Chinese adults. *Front Public Health.* (2021) 9:626331. doi: 10.3389/fpubh.2021.626331
40. Mueller SQ. Pre-and within-season attendance forecasting in Major League Baseball: a random forest approach. *Appl Econ.* (2020) 52:4512–28. doi: 10.1080/00036846.2020.1736502

41. Wang S, Wang Y, Wang D, Yin Y, Wang Y, Jin Y. An improved random forest-based rule extraction method for breast cancer diagnosis. *Appl Soft Comput.* (2020) 86:105941. doi: 10.1016/j.asoc.2019.105941

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

NOTATION

dc_i	i-th diagnosis code
$Ord(dc_i)$	Order of dc_i
$LCA(dc_i, dc_j)$	Least common ancestor of dc_i and dc_j
$s(dc_i, dc_j)$	Similarity of diagnosis code dc_i and dc_j
$S(D_i', D_j')$	Similarity of diagnostic information of patients i and j
$\mathbf{S}$	Patient similarity matrix based on diagnostic information
p_c	p coefficient to control the input exemplar preferences
K	Number of clusters
C_k	k-th cluster, $k = 1, 2,\ldots, K$
$E(C_k)$	Exemplar of cluster C_k
$Core_k$, $\lvert Core_k \rvert$	Core zone and the number of patients in C_k
$Prob_k\ (dc_h)$	Occurrence probability of the diagnosis code dc_h in C_k
$AOrd_k\ (Tdc_h)$	Average order of the typical diagnosis code dc_h in C_k
$Ord'\ (Tdc_h)$	New order of the typical diagnosis code dc_h
$TDCCoP_k$	k-th typical diagnosis code co-occurrence pattern
$LCoP_k$	k-th least common ancestor co-occurrence pattern
$\mathbf{CCoM}_k$	Conditional co-occurrence matrix for all diseases in $TDCoP_k$
UD_k	k-th unifying diagnosis
$IG(x_i)$	Information gain of feature x_i
$CVError_Z$	Average error using Z-fold cross-validation