Article published online: 2022-09-21

Conversion of Automated 12-Lead Electrocardiogram Interpretations to OMOP CDM Vocabulary

Sunho Choi[1,*] Hyung Joon Joo[2,3,*] Yoojoong Kim[4] Jong-Ho Kim[2,3] Junhee Seok[1]

[1] School of Electrical Engineering, Korea University, Seoul, South Korea
[2] Korea University Research Institute for Medical Bigdata Science, Korea University, Seoul, South Korea
[3] Department of Cardiology, Cardiovascular Center, Korea University College of Medicine, Seoul, South Korea
[4] School of Computer Science and Information Engineering, The Catholic University of Korea, Seoul, South Korea

Address for correspondence Junhee Seok, PhD, Korea University, 145, Anam-ro, Seongbuk-gu, Seoul 02841, South Korea (e-mail: jseok14@korea.ac.kr).

Appl Clin Inform 2022;13:880–890.

Abstract

Background A computerized 12-lead electrocardiogram (ECG) can automatically generate diagnostic statements, which are helpful for clinical purposes. Standardization is required for big data analysis when using ECG data generated by different interpretation algorithms. The common data model (CDM) is a standard schema designed to overcome heterogeneity between medical data. Diagnostic statements usually contain multiple CDM concepts and also include non-essential noise information, which should be removed during CDM conversion. Existing CDM conversion tools have several limitations, such as the requirement for manual validation, inability to extract multiple CDM concepts, and inadequate noise removal.

Objectives We aim to develop a fully automated text data conversion algorithm that overcomes limitations of existing tools and manual conversion.

Methods We used interpretations printed by 12-lead resting ECG tests from three different vendors: GE Medical Systems, Philips Medical Systems, and Nihon Kohden. For automatic mapping, we first constructed an ontology-lexicon of ECG interpretations. After clinical coding, an optimized tool for converting ECG interpretation to CDM terminology is developed using term-based text processing.

Results Using the ontology-lexicon, the cosine similarity-based algorithm and rule-based hierarchical algorithm showed comparable conversion accuracy (97.8 and 99.6%, respectively), while an integrated algorithm based on a heuristic approach, ECG2CDM, demonstrated superior performance (99.9%) for datasets from three major vendors.

Conclusion We developed a user-friendly software that runs the ECG2CDM algorithm that is easy to use even if the user is not familiar with CDM or medical terminology. We propose that automated algorithms can be helpful for further big data analysis with an integrated and standardized ECG dataset.

Keywords

- OMOP common data model
- ontology
- electrocardiogram
- clinical coding
- ECG interpretation
- ontology-lexicon

* *These authors equally contributed to the study.*

received
April 4, 2022
accepted after revision
July 29, 2022

DOI https://doi.org/10.1055/s-0042-1756427.
ISSN 1869-0327.

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

Background and Significance

The 12-lead electrocardiogram (ECG) is one of the most common tests performed in hospitals. It is used to screen and diagnose various cardiovascular diseases, including arrhythmia and coronary artery disease. Modern ECG machines can provide computerized ECG diagnostic statements that are comparable to physicians' interpretations.[1–4] Automated ECG interpretations can provide cost- and time-effective diagnostic decisions with less intra- and inter-observer variability. This tool may be very helpful for large-scale clinical studies, including population-based cohort studies, and in the future, personalized disease prediction. However, there are many ECG machine companies worldwide. Each company has its own ECG interpretation algorithm and unstandardized data structures for vocabulary and outputting results. Thus, standardizing the ECG result data must precede an integrated big data study using ECG data.[5–7]

The Observational Medical Outcomes Partnership (OMOP) is a public–private collaboration designed for methodological research to evaluate and establish scientific exemplary cases in observational data analysis.[8] The OMOP common data model (OMOP-CDM; CDM) is a standard data schema provided by the OMOP.[9] One of the main objectives of the CDM is to convert data in different formats, collected from various sources, into a common format with common representations of terminology, vocabulary, and coding schemes.[10] The power of the CDM is best illustrated in multicenter, big data research by reducing the heterogeneity between medical data.

Attempts to convert clinical data into CDM are currently underway in various fields.[11–14] It is essential to provide algorithms that transform medical data into a format applicable to the CDM. As advancements are made in the CDM, research on algorithms that transforms medical data into CDMs should be conducted in parallel. When using open-source software tools provided by the Observational Health Data Science and Informatics (OHDSI) such as WhiteRabbit, Rabbit-In-a-Hat, and USAGI, manual confirmation is required for connection and mapping of ECG result data to the field of a standardized glossary, provided by the CDM. Diagnostic statements usually contain multiple CDM concepts; however, in the case of the USAGI methodology, only one CDM concept can be recommended for each diagnostic statement. This can cause significant information loss and incorrect mapping, especially when a long non-essential information term is included in the diagnostic statement.

Objective

The aim of this study is to develop a fully automated algorithm for CDM-based conversion of 12-lead ECG interpretation text data that overcomes limitations of existing tools and manual conversion. To build an ontology-lexicon for an automation system, we analyzed the ECG result formats and vocabularies of three major ECG machine companies: GE Medical Systems Information Technology, Inc. (Milwaukee, Wisconsin, United States), Philips Medical Systems (Andover, Massachusetts, United States), and Nihon Kohden Co. (Tokyo, Japan). We named the mapping algorithm of the ECG dataset into the CDM database as ECG2CDM and evaluated its accuracy with that of the similarity comparison methodology and rule-based algorithm, which are representative methods commonly used in term-based approaches.

Methods

Data Acquisition

Here, 12-lead resting ECG tests were performed using automatic machines from three different vendors: GE Medical Systems, Philips Medical Systems, and Nihon Kohden. The ECG results were stored in the clinical information system (INFINITT Healthcare) of Korea University Anam Hospital in XML format from 2015. In total, 243,107 ECG results, from January 2019 to December 2019, were used in this study. After excluding duplicated statements, the number of statements included 139,704 from GE, 2,845 from Philips, and 153 from Nihon Kohden. The number of interpretations printed by these three vendors after removing duplicates was 142,506. This study involved the use of the hospital CDM database construction process. The study protocol was approved by the institutional review board of Korea University Anam Hospital (IRB NO. 2019AN0227). Written informed consent was waived by the institutional review board of Korea University Anam Hospital because of the retrospective study design that posed minimal risk to the participants. The study complied with the principles of the Declaration of Helsinki.

Ontology-Lexicon

The ontology-lexicon is defined by the existence of one or more lexica, specifying the concepts for every ontology.[15,16] As the number of terms is limited, an additional checkup is not required once the ontology-lexicon is properly implemented. It is also possible to solve the case shown in ►Table 1 when one statement term corresponds to several CDM terms. In this paper, a tool called ECG2CDM was created by utilizing cosine similarity and rule-based text processing using ontology-lexicon.

Cosine Similarity

In this study, we mapped the patient diagnostic statements to the corresponding pre-defined interpretation ontology-lexicon with the highest similarity.[16] For example, the diagnostic statement "BASELINE WANDER IN LEAD(S) V1,V2" can be mapped to the lexicon "baseline wander in lead(s) **." ►Fig. 1 shows an example of the mapping process of the ECG interpretation syntax using the similarity method.

Measuring similarity in text plays an important role in text-related research, such as information retrieval, text classification, and text condensation.[17] There are several pairwise similarity techniques. In this experiment, we used cosine similarity than other term-based similarity algorithms because accuracy is significantly important in

Table 1 Example of one-to-multiple correspondence between diagnostic statement and CDM vocabulary

Diagnostic statement	CDM ID	CDM concept name
Abnormal T, consider ischemia, anterior leads	4065390 4139185	EKG: T wave abnormal EKG: anterior ischemia
Abnormal T, consider ischemia, anterolateral leads	4065390 4139185 4137879	EKG: T wave abnormal EKG: anterior ischemia EKG: lateral ischemia
Atrial fibrillation or flutter	4064452 4065288	ECG: atrial fibrillation EKG: atrial flutter
Atrial fibrillation or flutter with aberrant conduction, or ventricular premature complexes	4064452 4065288 4089462	ECG: atrial fibrillation EKG: atrial flutter Ventricular premature complex
Atrial flutter with 2 to 1 block	4065288 4065290	EKG: atrial flutter ECG: partial atrioventricular block - 2:1
Biatrial enlargement	4100136 4101026	Left atrial abnormality Right atrial abnormality

Abbreviations: CDM, Common Data Model; EKG, electrocardiogram; ECG, electrocardiogram.

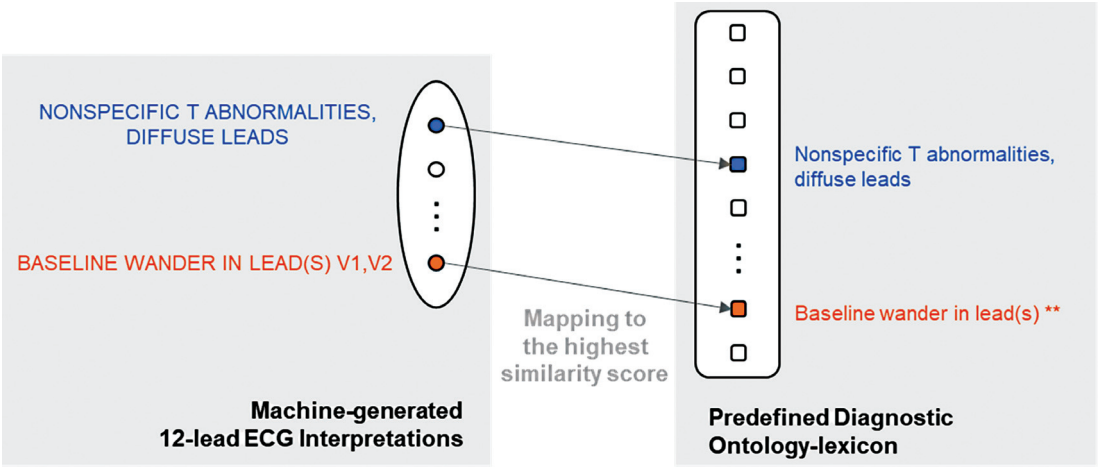

Fig. 1 Two examples of ECG interpretation syntax mapping by similarity. The first example completely matches the syntax existing in the ontology-lexicon with the patient's diagnosis, and the second example shows that the patient's diagnosis has the highest similarity to the syntax existing in the ontology-lexicon except for a few notes. ECG, electrocardiogram.

medical data.[18] Cosine similarity represents the similarity between two nonzero vectors of an inner product space. Each dimension in the vector represents a term that is a non-negative numeral. The cosine similarity is non-negative and bounded between 0 and 1. The two inputs become more similar as the similarity value approaches 1.

Rule-Based Text Processing

In cases of diagnostic statements with multiple possible ontologies, we attempted to extract all the ontology information in the diagnostic statement. For example, as shown in ►**Fig. 2**, the diagnostic term "atrial fibrillation with intermittent RBBB" can be regarded as the combination of two lexica "atrial fibrillation" and "RBBB."

Rule-based algorithms can be expressed in the form of a decision tree and flowchart, and the intuitive internal structure of the procedure makes it easier to understand than other algorithms.[19] In the rule-based text-processing algorithm, human modification is required to add new rules or change existing rules. If the rule is well organized without requiring changes, a rule-based system can perform well for overall problem-solving. In this study, the direction to set the rule was selected based on the designation of the type in the lexicon. By choosing this method, there is no need to change the existing rule even if new terms are added to the ontology-lexicon. We give each type to ontology-lexicon so that each term can be mapped while performing its own role. The detailed performing process is described in Result 4.1.

Tool Development

ECG2CDM, the algorithm to be presented in this paper, has been developed as a tool in two software formats. The standalone software was created in the form of an executable file using Python language and QT5. When producing web-based software, the web form was created using HTML and PHP, and the process of executing the algorithm was created using shell script and Python.

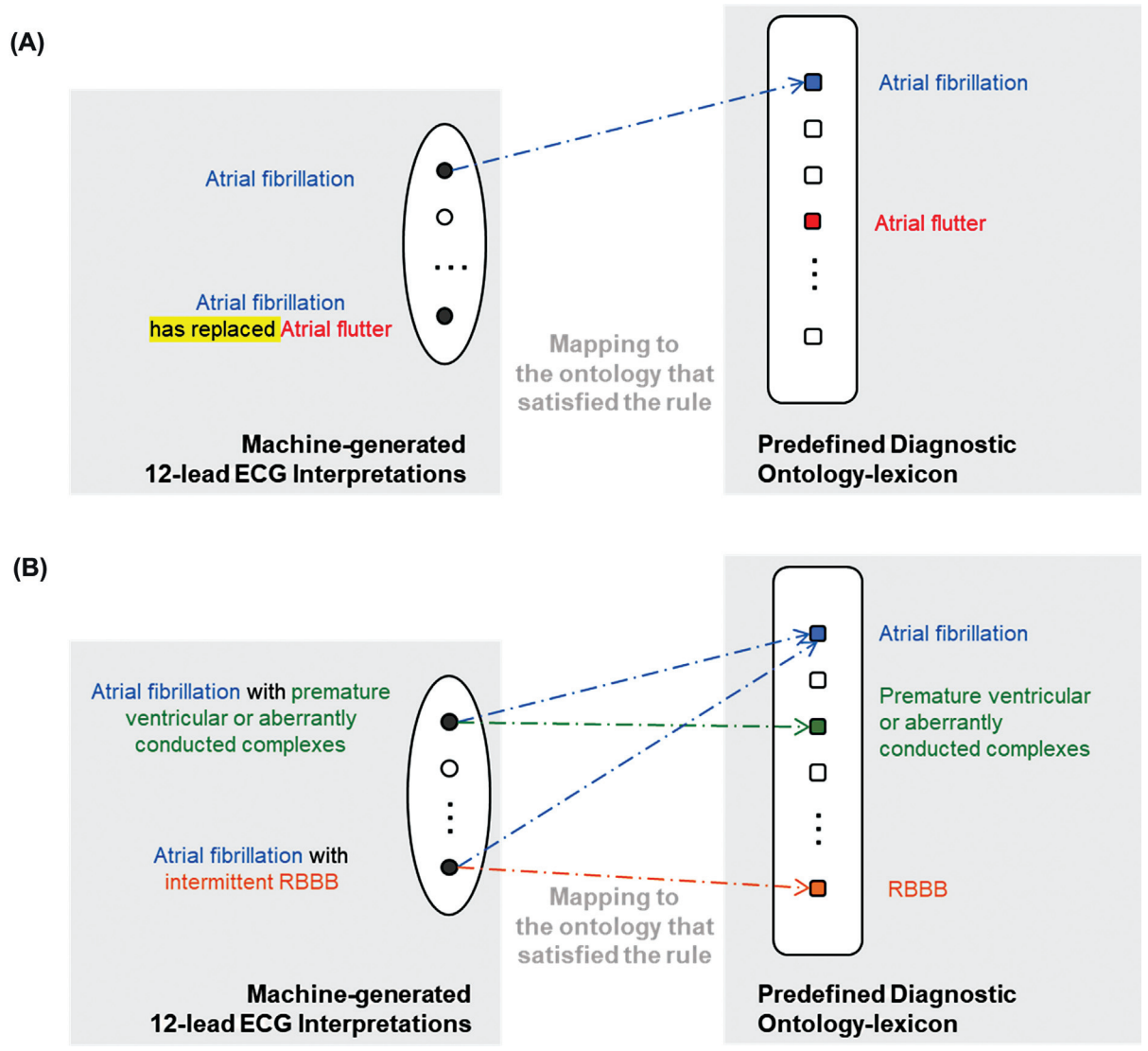

Fig. 2 Examples of an ECG interpretation syntax mapping in one-to-many correspondence. Panel **(A)** is an example of a rule in which the entire diagnosis is not mapped by the term "has replaced." In panel **(B)**, ECG interpretations generated by the rules of combining multiple ontologies have all the characteristics of each ontology. ECG, electrocardiogram.

As the stand-alone software was designed as a single executable file to run on a computer that does not have Python installed, it can only be executed in a Windows environment. The web-based software was designed for use in any operating system connected to the Internet. The ECG2CDM web-based software is supported by all major browsers. Both software are available at cdal.korea.ac.kr/ECG2CDM. If the environment of the computer is insufficient to apply both software, the user can modify the source code of ECG2CDM to be executable on the system. The Python code of the program is available at https://github.com/CDAL-SChoi/ECG2CDM. The anonymized patient dataset was also provided to test both software.

Results

In this study, we attempted to map machine-generated ECG interpretations to CDM concepts. The machine-generated ECG output contains more than one automatically generated diagnostic statement. Each statement can be mapped to one or more CDM code matching terms or may not be mapped to any CDM term. This study focused on the differences in the process of generating diagnostic statements. For automatic mapping, we first constructed an ontology-lexicon of ECG interpretations. We developed ECG2CDM, a combined diagnosis-mapping algorithm, using the ontology-lexicon.

Ontology-Lexicon of ECG Diagnostic Interpretations

To construct the ontology-lexicon, a medical coder labeled the diagnostic interpretations in the ECG machine user guidebooks of the three different vendors in accordance with the standard vocabulary of CDM, named ATHENA (athena.ohdsi.org). The ATHENA vocabulary includes various terms from international standards, such as SNOMED and ICDO. An experienced cardiologist performed secondary validation. Although there can be multiple CDM concepts in one interpretation, the interpretation may not map to any concept if there is no appropriate CDM concept or if the

Table 2 Example of type 1 diagnostic interpretations of GE mapped to the standard vocabulary of CDM

Interpretation term (Lexicon)	Type	CDM concept ID	CDM concept name
Aberrant complex	1	4092011	Aberrantly conducted complex
Aberrant conduction	1	4092011	Aberrantly conducted complex
Abnormal ECG	1	320536	Electrocardiogram abnormal
Abnormal left axis deviation	1	4215406	Left axis deviation
Abnormal QRS-T angle, consider primary T wave abnormality	1	4065390	EKG: T wave abnormal
Abnormal right axis deviation	1	4166218	Right axis deviation

Abbreviations: CDM, Common Data Model; ECG, electrocardiogram; EKG, electrocardiogram.

Table 3 Example of type 2, type 3, and type 4 diagnostic interpretations of GE mapped to the standard vocabulary of CDM

Interpretation term (Lexicon)	Type	CDM concept ID	CDM concept name
Are no longer	2		
Has changed	2		
PR interval has decreased	3	4171521	Shortened PR interval
PR interval has increased	3	4064885	Prolonged PR interval
Normal ECG	4	4065279	ECG normal
Normal rhythm	4	4142265	Normal sinus rhythm

Abbreviations: CDM, Common Data Model; ECG, electrocardiogram.

interpretation is non-essential information. In total, 593 ECG interpretations from Philips were mapped to 673 CDM standard vocabularies. In a similar process, 248 ECG interpretations from Nihon Kohden were mapped to 297 CDM standard vocabularies, and 453 ECG interpretations in GE were mapped to 340 CDM standard vocabularies. Through the ontology-lexicon structure, we implemented a case in which a single lexicon is mapped to several CDM concepts, or a case where the lexicon is not mapped to any CDM concept.

Actual diagnostic statements are constructed using lexica in ontology-lexicon and from physicians' notes that are not present in ontology-lexicon. Automatically generated diagnostic statements can be mapped to a single matching lexicon, to more than one lexicon, and clinically insignificant information such as name and date. Mapping with the CDM concepts corresponding to the ontology-lexicon proceeds when the diagnostic statement matches with a single lexicon. There are two ways to process mapping in the case of complex combinations of lexica and notes: mapping with the CDM concepts corresponding to all the included lexica or not conducting the mapping process, considering the entire information non-essential. To discern whether the diagnostic statement will be mapped, we divided the lexicon into various types.

The lexicon types are set to distinguish between the general lexicon that should be mapped and the exceptional lexicon that is idiomatically used. Lexica can be divided into four types; the most general form of the lexicon, type 1, can be output as a diagnostic statement on its own or in combination with other interpretations. For example, the lexicon "aberrant conduction," shown in ►**Table 2**, can be written as "aberrant conduction is now present" in an actual diagnostic statement and mapped to CDM concept "4092011 aberrantly conducted complex." However, combined with two lexica "atrial-sensed ventricular-paced rhythm" and "premature atrial complexes," the diagnostic statement becomes "atrial-sensed ventricular-paced rhythm with premature atrial complexes with aberrant conduction" and will be mapped to CDM concepts "4092011 aberrantly conducted complex," "4115173 atrial premature complex," and "4092041 ventricular pacing pattern."

As for other types, the lexica contain idiomatic expressions, as shown in ►**Table 3**. If the type 2 lexicon is included in the diagnostic statement, the entire statement is not mapped to any CDM concept. For example, the lexicon "are no longer" can be used as "borderline criteria for inferior infarct are no longer present" in the actual diagnostic statement. Since it is a record that the previous diagnosis no longer exists, it would not be mapped to any CDM concept. Type 3 lexica, specified for exception handling, contain type 2 terms but must be mapped to a CDM concept. Lastly, the type 4 lexicon is mapped only if the entire diagnostic statement matches it. Each type covers only 14, 11, and 3 lexica out of the total. The data of the constructed ontology-lexicon are available in ►**Supplementary Material Tables S1** to **S3** (available in the online version).

The relationship between the types should be established based on the order. This study used a rule-based hierarchical algorithm to solve complex combinations of lexica. Through the hierarchical configuration, it is possible to decide if mapping should be performed only when the conditions, according to the type, are met.

The pseudo-code of the rule-based hierarchical algorithm is as follows (►**Table 4**):

Table 4 Rule-based hierarchical algorithm

```
for Every diagnostic statement in diagnosis do
  if Type 2 lexicon is included in the diagnostic statement
  then
    if Type 3 lexicon is not included in diagnostic statement then
      Do not label any ontology
      Skip process to the next iteration
    else
      Continue the process
    end if
  end if
  if Diagnostic statement is fully matched with one of the Type 4
  lexicon then
    label as the fully matched ontology
  end if
  Search all ontology except Type 4 included in diagnostic
  statement
  Map diagnostic statement to all lexicon included
end for
```

ECG2CDM: A Standardization Algorithm

The number of words generated by each ECG machine is limited. The non-essential noise terms, which do not appear in the ECG machine user guidebook, are not labeled with CDM concepts. Therefore, it is important to filter the vocabulary through text preprocessing. In this study, the bag-of-words vectorizer was used to learn the tokens from the ontology-lexicon to remove unnecessary terms.[20] All ECG diagnostic statements were used in the ontologizing algorithm after converting words into the vector space using the preprocessed bag-of-words model. In detail, all possible terms that appear in the three ECG machine user guidebooks were collected in advance, the dimensions were configured as many as the number of terms, and embedding was performed by counting each term in the input statement. This process eliminates noise terms and unnecessary factors that can reduce the calculation speed and can cause confusion for clinicians to make accurate decisions.

In actual clinical data, where vendor information may be blank or incorrectly entered, it is important to analyze the term and execute the proper algorithm to map the ontology. The correspondence from the perspective of the term being created is easy to identify by discerning vendors, but conversely, it is not easy to predict the correspondence relationship by evaluating the term that has already been created. To solve this problem, we implemented a combined algorithm, ECG2CDM, as shown in ►Fig. 3.

The structure of ECG2CDM is designed to apply both similarity-based and rule-based algorithms. First, ECG2CDM checks whether the value of the result obtained by applying similarity to the input statement exceeds a certain threshold. When the value with the highest similarity is equal to or higher than the threshold when compared with the ontology, the result is mapped to the ontology with the highest similarity. Conversely, if the similarity value is less than the threshold, a rule-based algorithm is applied. At this time, there is a structural alteration from the pseudo-code of Algorithm 1, and the process of searching the full-text match with the term of type 4 is not included. This is omitted because the similarity is 1 when the full text is matched, and the mapping has already been performed for values greater than the threshold.

To specify the threshold using the heuristic method, the accuracy within the entire dataset was measured as the threshold changed. Consequently, as shown in ►Fig. 4, after a threshold value of 0.82, the accuracy was higher than 0.9975. Algorithms with a threshold of 0.82 or higher are considered significant in the study setting. In this experiment, we used an algorithm with a threshold of 0.90. The reason for choosing a threshold of 0.90 is that the accuracy shows a decreasing trend at threshold values greater than 0.91, and the robustness and the sensitivity can be reinforced by adjusting the ratio of the two algorithms.

Mapping Results

We conducted experiments with pairwise similarity and rule-based text processing, which are representative methods used in term-based text retrieval and classification, and compared their results with those of our algorithm. All models receive input embedded as a vector via bag-of-words. The performance was evaluated, in terms of accuracy, for the identical diagnostic statement dataset obtained from each vendor and from the vendors collectively as each vendor has different interpretation algorithms and different amounts of diagnostic statement data. To ensure the validity of the results, these experiments were repeated 100 epochs to measure the performances by randomly extracting 80% of the data from the entire dataset. We evaluated the average and confidence interval of accuracy.

To check the effectiveness of this structure and determine the scope for improvement, experiments were conducted using the deduplicated ECG diagnostic statement data. ►Table 5 shows the mean values and 95% confidence interval results in 100 experiments of two representative algorithms and ECG2CDM. In all experiments, ECG2CDM showed superior performance to the other two algorithms.

Conversion rates were also compared when each algorithm was applied to the entire dataset. The cosine similarity has an accuracy of approximately 0.9778 over the entire dataset, with 139,390 of the diagnostic statements correctly mapped. The results of the extant rule-based algorithm matched 142,126 cases. Compared with the other methods, ECG2CDM achieved the highest accuracy on the same dataset. Using the ECG2CDM algorithm, with a 0.90 threshold, 142,422 diagnostic sentences were correctly mapped, which represent an accuracy of approximately 0.9994 in the total dataset. In other words, more than 320 cases were accurately predicted compared with the results of the existing algorithms, and data other than 85 cases were correctly predicted.

Standardization Software Tools

Herein, we present two types of demonstrations of the ECG2CDM software, as shown in ►Fig. 5. The first is the use of the ECG2CDM algorithm for a single diagnostic

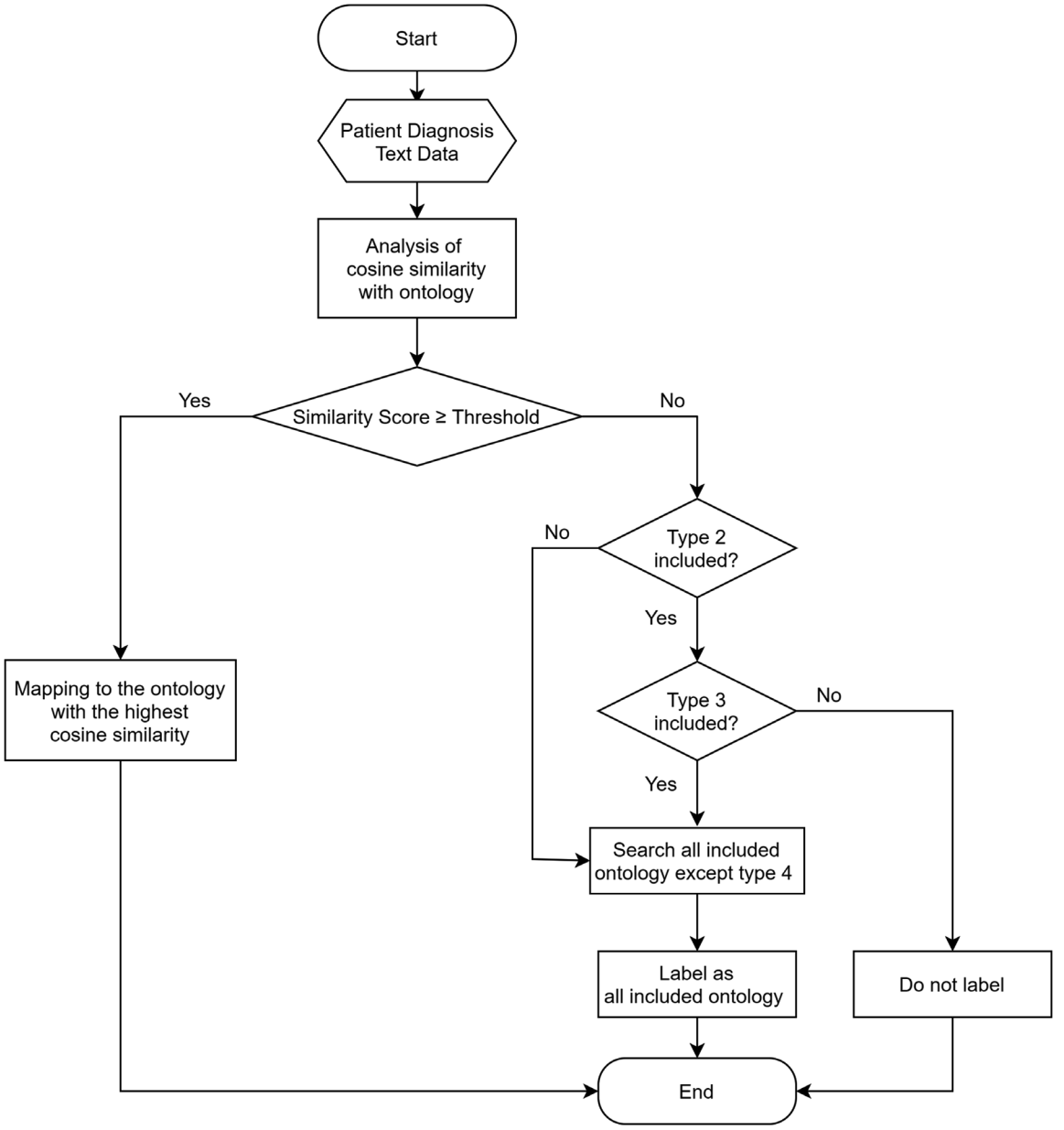

Fig. 3 Flowchart of the ECG2CDM algorithm. The structure of ECG2CDM is designed to apply both similarity-based and rule-based algorithms. When the similarity score obtained from the cosine similarity is above a certain threshold, ECG2CDM emits the result of the similarity-based algorithm. For the other case, ECG2CDM emits the result of the rule-based algorithm.

statement. The software outputs CDM terms that can be converted when the user enters a diagnostic statement printed on an actual 12-lead resting ECG test or written by medical personnel. The user can run the task by directly entering a single-term input into a text box. After entering the text and submitting it, the information, converted to the CDM code, is provided in a tabular form. This tool is provided in the ECG2CDM web-based software.

In the second demonstration, when loading the file in which diagnostic statements are stored, the file with information converted to ontology is output. The user must upload a file in xlsx or csv format that stores diagnostic statements in the form of a list in the "statement" column. The algorithm then outputs a downloadable csv file when the input is submitted. This tool is available in the ECG2CDM web-based software and stand-alone software.

In the case of web-based software with two demonstrations, diagnostic statement data or file data are temporarily saved and then ECG2CDM python code is executed on a separate server. Data that may contain personal information is removed after execution, leaving only dummy data in the initial state before data are input into the server. Accordingly, data that may cause privacy issues will perish after execution.

Discussion

Electronic health record analysis using an automated algorithm is more advantageous in finding a specific diagnosis or

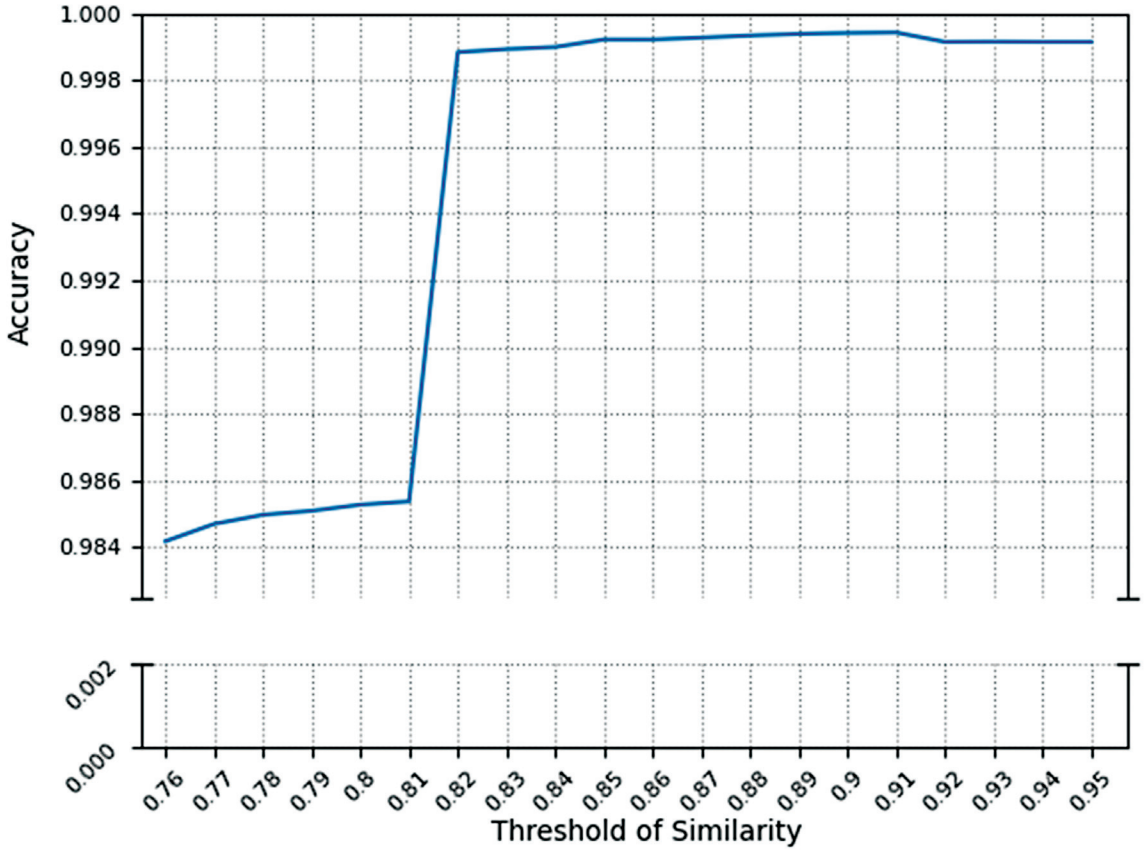

Fig. 4 Variation in the accuracy of ECG2CDM algorithm according to the threshold. The result shows an upward trend as the threshold increases, but at a threshold of 0.9 or higher, it slightly decreases and shows a tendency to remain as it is.

Table 5 Conversion accuracy for the collected automated 12-lead ECG interpretation dataset

Dataset	Methods	Average Accuracy	(95% Confidence Interval)
Total ($n = 142,506$)	Cosine Similarity	0.9781	(0.9780, 0.9782)
	Rule-based Similarity	0.9960	(0.9960, 0.9960)
	ECG2CDM	**0.9994**	**(0.9994, 0.9994)**
GE ($n = 139,704$)	Cosine Similarity	0.9809	(0.9809, 0.9810)
	Rule-based Similarity	0.9988	(0.9988, 0.9988)
	ECG2CDM	**0.9994**	**(0.9994, 0.9994)**
Philips ($n = 2,845$)	Cosine Similarity	0.8234	(0.8227, 0.8242)
	Rule-based Similarity	0.8813	(0.8807, 0.8819)
	ECG2CDM	**0.9972**	**(0.9971, 0.9973)**
Nihon Kohden ($n = 153$)	Cosine Similarity	0.9930	(0.9924, 0.9936)
	Rule-based Similarity	0.5575	(0.5536, 0.5615)
	ECG2CDM	**0.9935**	**(0.9929, 0.9942)**

Abbreviations: CDM, Common Data Model; ECG, electrocardiogram.

standardizing data compared with manual coding-based analysis.[2,21] Currently, research on automated clinical coding and software development is being actively conducted in various medical fields.[22,23] There has been an attempt to construct a standard ontology specialized for ECG,[24] but an automated mapping tool has not been developed in relation to ECG diagnosis. In this study, we developed an automated CDM mapping algorithm called ECG2CDM, based on similarity-based and rule-based hierarchical algorithms. Using our tool, diagnostic statements, generated through 12-lead resting ECG tests, can be converted into CDM codes without the need for manual checks or secondary confirmations. We

present an algorithm with a conversion rate of approximately 0.9994 on 142,422 out of 142,506 cases. We believe that our research can contribute to increasing the interoperability of ECG data by standardizing the interpretation, using the global standard vocabulary.

Using the ECG2CDM algorithm presented in this paper, not only the anonymized ECG test but also the database used in the actual hospital was converted. It is applied in the system used by actual clinicians. All ECG test data collected from 868,981 patients from 2013 to 2018 in Anam Hospital have been converted, and we are attempting to switch for all data. This allows the use of analysis tools such as ACHILLES and ATLAS provided by OHDSI and enables cooperation with other hospital systems.

Of the 84 cases where the mapping was not accurate, those with the highest percentage are the cases mapped to a broad range of terms rather than detailed. For example, out of 37 cases, "ST elevation in Inferior leads" was not mapped to "inferior ST segment elevation" but converted to "ST segment elevation," and thus the information was abbreviated. Furthermore, there were 17 cases that caused problems such as "accelerated junctional rhythm" mapping to itself as well as "junctional rhythm." When adjectives are added, they are double mapped to the original text as well as the term that refers to the whole. In other incorrectly mapped cases, a diagnostic sentence composed of terms that are not in the ontology-lexicon may be entered, and some information may be omitted. We considered that inadequate mappings in these cases did not result in significant information loss or misinterpretation.

The experimental setting in this paper compares the ontology-lexicon with the entered statement and finds the one with the best similarity or included in the statement. This might cause a problem that may not properly extract the meaning of the statement in the context of natural language. However, this study analyzed semi-standardized ECG test output from an ECG machine, a different approach was used from other natural language processing (NLP). In the ECG test data collected from three devices, it is expressed using type 2 terms as in "premature atrial complexes are no longer present" and "ST no longer depressed in anterior leads" when describing "symptom disappears," or is expressed using type 4 term when describing "no issue appeared." From this point of view, all cases that could lead to syntactical misleading, as identified so far for the ECG test data, were processed normally. We expect that input data newly generated from the three vendors will not deviate from this format.

We demonstrated that the performance of our algorithm surpasses that of conventional methodologies. We suggest several hypotheses for performance improvement in ECG2CDM. First, the algorithm becomes less sensitive to secondary noise information, such as comments, notes, and terms that are not used in the actual mapping by setting a threshold. In other words, if only the score of 0.90 or less of the terms resemble a certain ontology when the process is performed for all ontologies, we determined that the part that matches the ontology acts as an important factor, not a secondary term. We believe that the robustness of the optimal algorithm was reinforced by setting the threshold.[25]

Second, we were able to minimize the inaccuracy in CDM mapping process by combining two algorithms. In the case of Nihon Kohden, CDM mapping was easily achieved by a simple similarity algorithm because the output diagnostic statements were in the form of affixed comments to a single ontology. This can be taken for granted because the ontology-lexicon composed of terms extracted from Kohden has only type 1 term, which is provided in supplementary data. In the cases of Philips and GE, CDM mapping was often a one-to-many correspondence problem, which required a more complex rule-based algorithm because the output diagnostic statement was generated in the form of multiple combinations of ontologies. Ideally, it would be possible to apply an appropriate algorithm to each vendor. In reality, some patients were often tested multiple times by different vendor ECG machines, which considerably decreased the performance of the single correspondence mapping strategy (►Table 5).

The ontology used in the ECG2CDM was created using ECG interpretation outputs described in the vendor guidebooks. These guidebooks can be revised on a regular basis during device upgrades or through the extension of diagnosis;

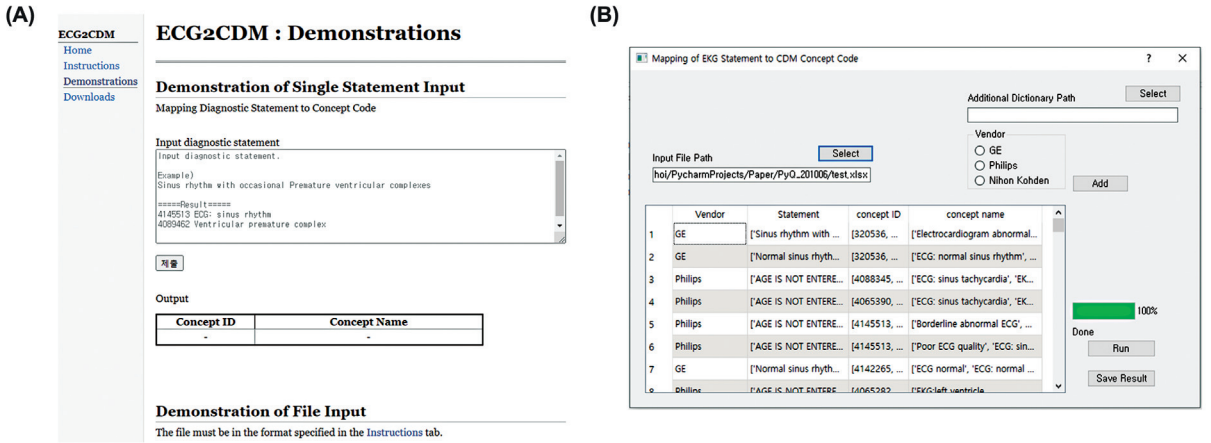

Fig. 5 ECG2CDM software interfaces provided as a **(A)** web-based software and **(B)** stand-alone software. Both software accepts statements from the ECG test as input. ECG, electrocardiogram.

however, the diagnostic statement output algorithm will not change significantly. Therefore, a minor upgrade of ontology-lexicon will be sufficient to maintain ECG2CDM performance even after vendor upgrades, rather than changing the inner structure of ECG2CDM.

To demonstrate the general applicability of ECG2CDM, 44 primary standardized diagnostic statements identifiable in the Shandong Provincial Hospital 12-lead ECG dataset[26] were converted into CDM concepts by ECG2CDM. The data of the converted primary statements are available in ►**Supplementary Table S4** (available in the online version). There were 43 cases of CDM concepts corresponding to a primary statement except for the statement "ST-T change due to ventricular hypertrophy." Among them, 34 cases were correctly mapped to CDM concepts through ECG2CDM, and the conversion rate for statements non-generated on three devices was around 0.79. Most of the cases not mapped correctly were because the dataset used a word or expression that is not in the existing dictionary, such as abbreviating "myocardial infarction" to "MI." This demonstration shows the applicability of ECG2CDM to general statements when terms in a dictionary are used.

In terms of specificity, as a result of extracting the mapping results for 192 statements that appeared only once, excluding date information, in the dataset from the converted result of ECG2CDM, a total of 187 cases were converted correctly. This analysis showed that ECG2CDM was able to convert even for infrequent statements with an accuracy of 0.97.

NLP systems are known to address a wide variety of important clinical and research tasks compared with statistical methods.[27] Although our study did not use NLP systems, it was shown that statistical and rule-based hierarchical algorithms can achieve remarkable mapping accuracy in the case of ECG. For complex problems, rule-based and probabilistic methodologies should be used, and machine learning and deep learning should be combined for performance improvement as NLP is useful for unstructured reports and free-form reports.[27,28] We expect that our study will be helpful in the application of CDM research in various medical fields and in complex data processing.

Conclusion

The proposed ECG2CDM algorithm for an automated 12-lead ECG dataset showed a superior performance gap compared with conventional text-processing methods. Furthermore, we developed user-friendly software.

Clinical Relevance Statement

Using an automatic text data conversion algorithm specialized for 12-lead ECG diagnostic statements can facilitate mapping the semi-standardized ECG test data into the OMOP CDM-based standardized vocabulary. By refining data collected from various devices, it is possible to minimize confusion caused by inconsistent sentences, enabling clinicians to make accurate decisions. Furthermore, converted multicenter medical data can accelerate related research since the problem of data heterogeneity and anonymization has been solved.

Multiple Choice Questions

1. Which of the following refers to the standard vocabulary provided by OHDSI?
   a. USAGI.
   b. ATHENA.
   c. WhiteRabbit.
   d. Rabbit-In-a-Hat.

   Correct Answer: The correct answer is option b. All options are provided by OHDSI, but all other options except for option B are tools for designing ETL (WhiteRabbit, Rabbit-In-a-Hat) and mapping vocabulary (USAGI).

2. Which of the following is a characteristic of cosine similarity?
   a. The similarity value is determined by subtraction between two inputs.
   b. The two inputs become more similar as the similarity value approaches 0.
   c. The similarity value is an integer type.
   d. The similarity value is nonnegative.

   Correct Answer: The correct answer is option d. The similarity value is determined by an inner product between two inputs. The two inputs become more similar as the similarity value approaches 1, and the similarity value is a float type bounded between 0 and 1.

Protection of Human and Animal Subjects

The study protocol was approved by the institutional review board of Korea University Anam Hospital (IRB NO. 2019AN0227). Written informed consent was waived by the institutional review board of Korea University Anam Hospital because of the retrospective study design that posed minimal risk to the participants. The study complied with the principles of the Declaration of Helsinki.

Funding

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI19C0360) as well as a grant from the National Research Foundation of Korea (grant number: NRF-2022R1A2C2004003).

Conflict of Interest

None declared.

References

1 Garcia TB. 12-Lead ECG: The Art of Interpretation. Jones & Bartlett Publishers; 2013

2 Smulyan H. The computerized ECG: friend and foe. Am J Med 2019;132(02):153–160
3 Willems JL, Abreu-Lima C, Arnaud P, et al. The diagnostic performance of computer programs for the interpretation of electrocardiograms. N Engl J Med 1991;325(25):1767–1773
4 Kligfield P, Gettes LS, Bailey JJ, et al; American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; American College of Cardiology Foundation; Heart Rhythm Society. Recommendations for the standardization and interpretation of the electrocardiogram: part I: the electrocardiogram and its technology a scientific statement from the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society endorsed by the International Society for Computerized Electrocardiology. J Am Coll Cardiol 2007;49(10):1109–1127
5 Reich C, Ryan PB, Stang PE, Rocca M. Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases. J Biomed Inform 2012;45(04):689–696
6 Gonçalves B, Guizzardi G, Pereira Filho JG. Using an ECG reference ontology for semantic interoperability of ECG data. J Biomed Inform 2011;44(01):126–136
7 Khan U, Kothari H, Kuchekar A, Koshy R. Common Data Model for Healthcare Data. Paper presented at: 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information \& Communication Technology (RTEICT); 2018:1450–1457
8 Stang PE, Ryan PB, Racoosin JA, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. Ann Intern Med 2010;153(09):600–606
9 Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc 2012;19(01):54–60
10 Makadia R, Ryan PB. Transforming the Premier Perspective Hospital Database into the Observational Medical Outcomes Partnership (OMOP) Common Data Model. EGEMS (Wash DC) 2014;2(01):1110
11 Sathappan SMK, Jeon YS, Dang TK, et al. Transformation of electronic health records and questionnaire data to OMOP CDM: a feasibility study using SG_T2DM dataset. Appl Clin Inform 2021;12(04):757–767
12 Lamer A, Depas N, Doutreligne M, et al. Transforming French Electronic Health Records into the observational medical outcome partnership's common data model: a feasibility study. Appl Clin Inform 2020;11(01):13–22
13 Lynch KE, Deppen SA, DuVall SL, et al. Incrementally transforming electronic medical records into the observational medical outcomes partnership common data model: a multidimensional quality assurance approach. Appl Clin Inform 2019;10(05):794–803
14 Maier C, Lang L, Storf H, et al. Towards Implementation of OMOP in a German University Hospital Consortium. Appl Clin Inform 2018;9(01):54–61
15 Cimiano P, McCrae J, Buitelaar P, Montiel-Ponsoda E. On the role of senses in the ontology-lexicon. In: New Trends of Research in Ontologies and Lexical Resources. Springer; 2013:43–62
16 Cimiano P, Unger C, McCrae J. Ontology-based interpretation of natural language. Synth Lect Hum Lang Technol. 2014;7(02):1–178
17 Gomaa WH, Fahmy AA. others. A survey of text similarity approaches. Int J Comput Appl 2013;68(13):13–18
18 Hogan WR, Wagner MM. Accuracy of data in computer-based patient records. J Am Med Inform Assoc 1997;4(05):342–355
19 Sasaki M, Kita K. Rule-based text categorization using hierarchical categories. Paper presented at: SMC'98 Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 98CH36218). Vol 3.; 1998:2827–2830
20 Zhang Y, Jin R, Zhou Z-H. Understanding bag-of-words model: a statistical framework. Int J Mach Learn Cybern 2010;1(1–4):43–52
21 Wadia R, Akgun K, Brandt C, et al. Comparison of natural language processing and manual coding for the identification of cross-sectional imaging reports suspicious for lung cancer. JCO Clin Cancer Inform 2018;2:1–7
22 Catling F, Spithourakis GP, Riedel S. Towards automated clinical coding. Int J Med Inform 2018;120:50–61
23 Ternois I, Escudié J-B, Benamouzig R, Duclos C. Development of an Automatic Coding System for Digestive Endoscopies. EFMI-STC; 2018:107–111
24 Zouri M, Zouri N, Ferworn A. An Ontology Approach for Knowledge Representation of ECG Data. ITCH; 2019:520–525
25 Agassounon W, Martinoli A. Efficiency and robustness of threshold-based distributed allocation algorithms in multi-agent systems. Paper presented at: Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: Part 3.; 2002:1090–1097
26 Liu H, Chen D, Chen D, et al. A large-scale multi-label 12-lead electrocardiogram database with standardized diagnostic statements. Sci Data 2022;9(01):272
27 Kreimeyer K, Foster M, Pandey A, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. J Biomed Inform 2017;73:14–29
28 Kim Y, Lee JH, Choi S, et al. Validation of deep learning natural language processing algorithm for keyword extraction from pathology reports in electronic health records. Sci Rep 2020;10(01):20265