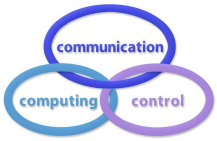
INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 19, Issue: 1, Month: February, Year: 2024
Article Number: 6251, https://doi.org/10.15837/ijccc.2024.1.6251



CCC Publications

Automated Generation of ICD-11 Cluster Codes for Precision Medical Record Classification

Jiayi Feng, Runtong Zhang, Donghua Chen, Lei Shi, and Zhaoxing Li

Jiayi Feng
Department of Information Management
Beijing Jiaotong University, China
No.3, Shangyuan Village, Haidian District, Beijing 100044, China
jyfeng@bjtu.edu.cn

Runtong Zhang*
Department of Information Management
Beijing Jiaotong University, China
No.3, Shangyuan Village, Haidian District, Beijing 100044, China
*Corresponding author:rtzhang@bjtu.edu.cn

Donghua Chen
Department of Information Management
University of International Business and Economics, China
No.3, Shangyuan Village, Haidian District, Beijing 100044, China
dhchen@uibe.edu.cn

Lei Shi
School of Computing
Newcastle University, UK
Open Lab, Floor 1 Urban Sciences Building, Newcastle upon Tyne NE4 5TG, United Kingdom Open Lab, NE4 5TG, Newcastle upon Tyne, United Kingdom
lei.shi@ncl.ac.uk

Zhaoxing Li
Department of Electronics and Computer Science
University of Southampton, UK
B32, East Highfield Campus, University Road SO17 1BJ, Southampton, United Kingdom
zhaoxing.li0808@outlook.com

Abstract

Accurate clinical coding using the International Classification of Diseases (ICD) standard is essential for healthcare analytics. ICD-11 introduces new coding guidelines and cluster structures, posing challenges for existing coding tools. This research presents an automated approach to generate valid ICD-11 cluster codes from medical text. Natural language records are represented as vectors and compared to an ICD-11 corpus using cosine similarity. A bidirectional matching technique then refines similarity estimation. Experiments demonstrate the method yields up to 0.91 F1 score in coding accuracy, significantly outperforming a baseline tool. This work enables efficient high-quality ICD-11 coding to support healthcare informatics.

Keywords: ICD-11, ICD code, machine learning, text similarity, clinical coding.

1 Introduction

Health care is one of the most important areas of social development and well-being[1]. Encoded health data is crucial for healthcare service financing, physician remuneration, and medical research[2]. Diagnoses recorded in electronic health records are identified using the International Classification of Diseases (ICD) codes[3]. The accuracy of disease-related groups derived from the ICD coding system is determined by the quality of ICD coding[4]. Due to the digital revolution, the amount of data that needs to be processed is growing every day[5]. Correspondingly, there has been a growing demand for improved medical record coding quality based on ICD. Most hospitals still follow traditional disease coding procedures, wherein after a patient is discharged, the medical record is archived and then coded manually or semi-automatically by the coder using the ICD coding principles and terminology dictionary. The current scenario of assigning clinical codes is a highly expensive, time-consuming, and error-prone manual process[6]. In addition, ICD coding is known to be complex and difficult, and the results obtained from it vary greatly depending on the skill level and proficiency of each in-dividual coder[7]. Computers and automation can enable and inspire new ways of working[8], and computer-assisted systems can significantly improve the efficiency of ICD coding and reduce labor waste[9].

Existing computer-assisted coding tools, such as the online ICD coding tool provided by the World Health Organization and the publicly available disease query system offered by MedSci[10], support inputting query keywords to retrieve relevant disease names and codes. However, in practice, simple keyword-based retrieval often encounters issues of missing terms. Rule-based approaches[11–13] or machine learning models with manual features[14–17] have been utilized to address this problem by extracting keywords from electronic medical records using natural language processing techniques, classifying diseases, and resolving polysemy. Nevertheless, these methods still rely on dictionary mapping for coding. Additionally, deep learning models[18, 19] have been employed to address the aforementioned issues by modeling large amounts of historical coding data. However, the complexity of application coding rules poses challenges for such methods, and the requirement for a substantial volume of historical coding data makes it less feasible for early adoption of new disease classification standards.

Furthermore, the significant differences between the 11th Revision of the International Classification of Diseases (ICD-11) and other ICD versions make upgrading existing coding systems a difficult and time-consuming task. ICD-11 was officially released during the World Health Assembly in 2019, marking the latest chapter in the evolution of this globally adopted classification system for diseases[20]. With each new iteration of the ICD, it is customary for the code syntax to undergo modifications, ostensibly to avert confusion with previous versions[21]. The ICD-11 introduced new coding structures and rules, including Cluster Coding[22]. As a result, it is difficult to use existing coding systems or outdated ICD coding systems[23] to generate multiple codes that are suitable for cluster coding under the new ICD-11 classification standard.

For healthcare services that generate large amounts of data, machine learning (ML) has proven to be a useful tool to aid decision-making[24, 25]. Therefore, it is essential to establish an automated coding methodology for ICD-11 cluster coding based on existing automated coding techniques and ML methods, simplified ICD coding business processes, which could improve the quality of service operations management in hospitals[26], and improve the accuracy and efficiency of computer-assisted ICD-11 coding as shown in Figure 1.

This study proposes a method for disease classification based on automatic generation of ICD-11 codes in medical record. The proposed method amalgamates the coding framework and rules of ICD-11, while building a model that maps the natural language narrative of electronic medical records to ICD-11 codes. The research in this paper includes three main contributions. Firstly, the proposed ICD-11 coding model is based on the coding rules of ICD-11 and is intended to overcome the limitation of current ICD coding tools in generating ICD-11 cluster coding automatically. Secondly, a bidirectional matching model for disease diagnosis text similarity calculation was established by combining text similarity and feature word weights, to refine disease text similarity calculation. Finally, an ICD-11 coding generation system based on medical records was proposed to facilitate the automatic generation of clustered coding and ensure the accuracy and completeness of the resulting clustered codes.

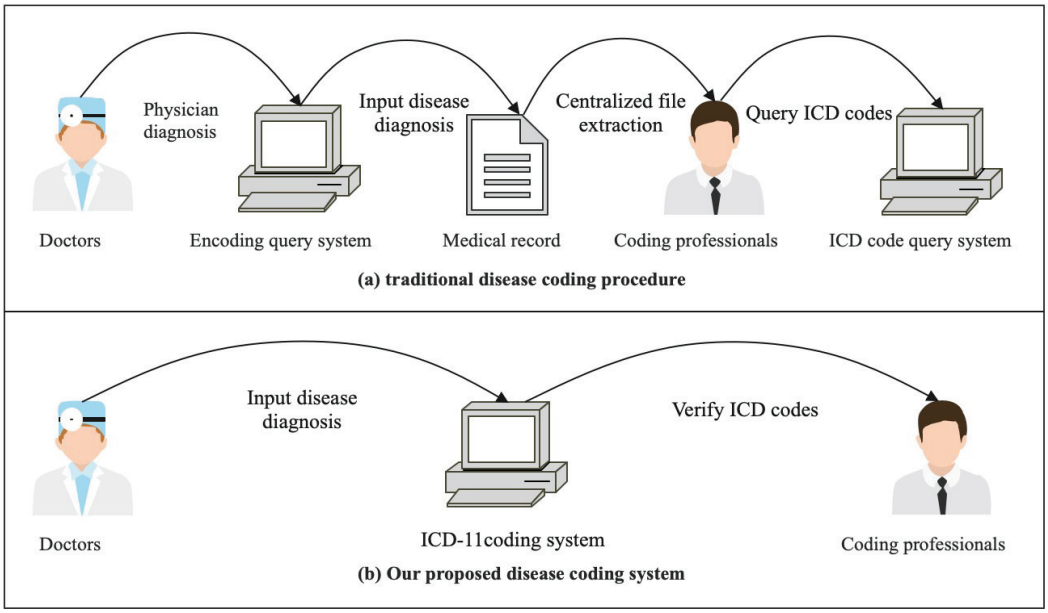


Figure 1: Comparison of disease coding process of the traditional method and our proposed method

The materials and methods section presents the ICD-11 cluster coding model and the bidirectional matching-based similarity correction algorithm. The results section presents the results of applying the methodology to a dataset of medical records, including performance metrics and comparisons with existing methods. The discussion section analyzes the findings, addresses limitations, and highlights the practical applications of the approach. Finally, the conclusion summarizes the contributions of the paper and discusses avenues for future research.

2 Materials and Methods

2.1 ICD-11 cluster coding model

ICD-11 represents a significant milestone in the evolution of healthcare information standards, with profound implications for the advancement of healthcare informatics in the decades to come [27]. Compared to the ICD-10 standard, the ICD-11 has adopted a new coding framework, resulting in changes to its coding format. The ICD-11 employs a combination coding system, consisting of Stem Code and Extension Code. The Stem Code (S) represents the primary code and defines the fundamental classification framework, while the Extension Code (E) represents the supplementary code, providing detailed additional information. It is important to note that the Extension Code cannot be used independently and must be used in conjunction with the Stem Code. A Stem Code can be combined with multiple Extension Codes. Additionally, not all Extension Codes can be matched with any Stem Code. For example, if a patient is diagnosed with GC08.0 Urinary tract infection caused by Escherichia coli, unspecified site with the characteristic of "MG50.27 Escherichia coli producing broad-spectrum β-lactamase", then the ICD-11 coding system needs to generate the combined code "GC08.0/MG50.27". The universal form of cluster coding CC can be defined as

$$CC = S_1\&E_{1,1}\&E_{1,2}/S_2\&E_{2,1}/\cdots/S_n, \tag{1}$$

where $E_{1,1}$ represents the first Extension Code associated with S_1, n represents the number of Extension Codes, "&" symbol denotes the connector between Stem and Extension Codes, and "/" symbol indicates the relationship between two Stem Codes.

Stem Codes can also include all information in a pre-combined form, known as pre-coordination. For example, the ICD-11 code "2C25.2 Squamous cell carcinoma of bronchus or lung" incorporates both the body site and pathology in a pre-coordinated Stem Code. The Extension Codes in equation (1) are used based on the Sanctioning Tables, which define three related scopes of use: mandatory, permitted, and not permitted.

To ensure the accuracy of ICD-11 code generation, we strictly observe the ICD-11 encoding principles during the implementation of the corresponding encoding query and generation functions in the ICD-11 cluster coding model. We present a method for automatically coding medical records based on the ICD-11 classification system. The proposed approach involves a computer-assisted automatic coding method for ICD-11, which relies on trusted corrected similarity scores. We establish a bidirectional matching similarity correction algorithm based on term frequency-inverse document frequency (TF-IDF)[28] and cosine similarity[29] using the ICD-11 basic data, which includes the ICD-11-MMS version published by the World Health Organization. Figure 2 illustrates the flowchart of the proposed ICD-11 automatic coding method.

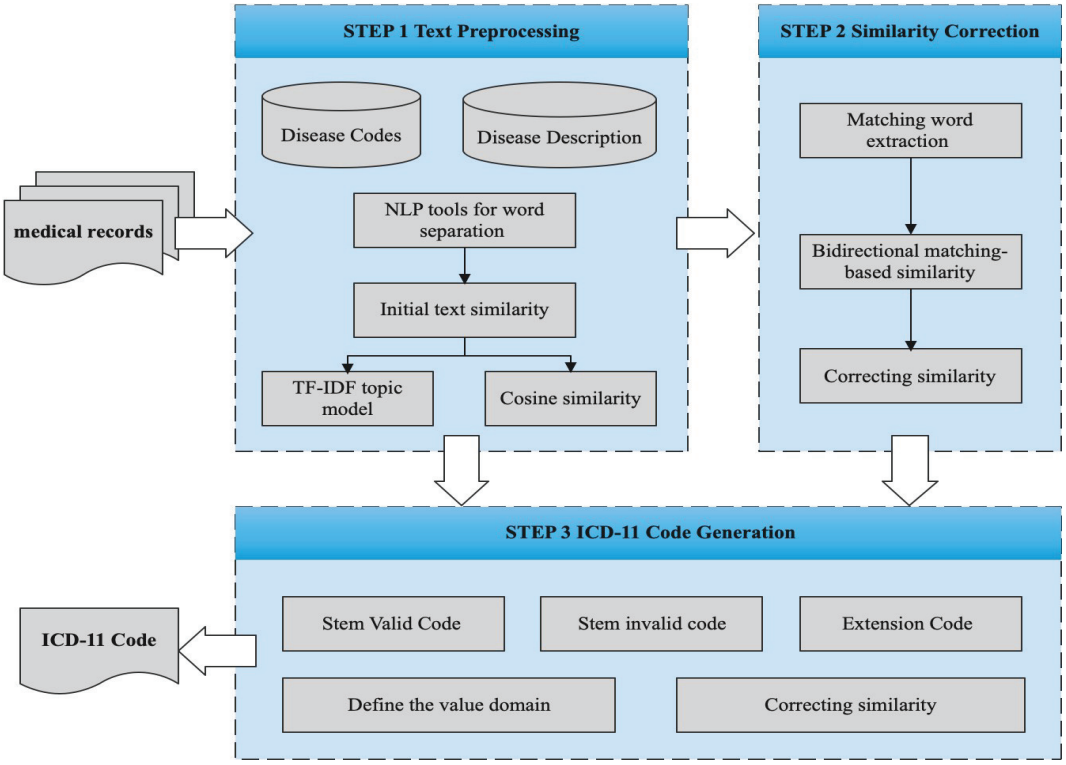


Figure 2: Flowchart of computer-automated coding based on text similarity estimation

2.2 Disease diagnosis text similarity calculation model

This study proposes a hybrid method, which combines string and corpus-based similarity calculations. String-based similarity calculations are applied directly to the raw text and evaluate similarity in terms of character matching or distance. The corpus-based method handles the corpus, which is a large collection of written or spoken text often used for language studies. Language corpora used for semantic similarity calculations are specific to the task domain and can differ in their selection criteria. In the case of the ICD-11 coding corpus, the deviation between diagnostic text by medical professionals and the coding corpus is negligible. Using this corpus, we can transform the text into a semantic vector representation, calculate similarities with text in electronic medical records, and obtain the highest similarity score for a specific code combination, which we can then utilize to map the text to the corresponding ICD-11 code. We apply the vector space model from the string-based method to the vector representation form of text[30].

As in equation (2), the cosine similarity algorithm considers the word frequency and global frequency of the text itself and belongs to a type of similarity algorithm based on the vector space model. In the generation of ICD-11 codes, natural language medical records are transformed into vector representations and compared to the vector representations of the ICD coding corpus.

$$\text{cosine}\,\theta = \frac{a-b}{\|a\|\bullet\|b\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{2}$$

Due to the special characteristics of electronic medical record data, conventional processing methods prove inadequate in extracting feature terms from medical narratives elucidated by other practitioners employing natural language. Nevertheless, the TF-IDF algorithm has ascended as an indispensable tool in the realm of information retrieval. By leveraging a mathematical formula, the TF-IDF topic modeling algorithm determines the weight of each term in an EMR, based upon its frequency in the document and its relative frequency in the medical corpus. Employing a weighting scheme that assigns importance to terms that are prevalent in the EMR yet arise infrequently in the corpus, the TF-IDF algorithm facilitates the identification and extraction of medically consequential features from the medical records. The equations (3)-(5) below demonstrate the calculation of TF-IDF.

$$TF(i,j) = \frac{N_{i,j}}{\sum_k N_{k,j}} \tag{3}$$

$$IDF(j) = \log \frac{|D|}{|j : t_i \in d_j| + 1} \tag{4}$$

$$TF - IDF = TF * IDF \tag{5}$$

where $\text{TF}(i,j)$ refers to the frequency that term i appears in document j, and $IDF(i)$ refers to the inverse document frequency for term i. $\sum_k N_{k,j}$ represents the sum of all word occurrences within d_i, while $|D|$ refers to the total number of documents within the corpus.

2.3 The bidirectional matching-based similarity correction algorithm

The similarity correction algorithm based on bidirectional matching is a refinement of traditional text similarity algorithms, which incorporates the unique features of ICD-11 disease description texts. While the initial similarity scores derived from the TF-IDF topic model and cosine text similarity algorithms provide a reasonable foundation upon which to build, they cannot fully and accurately reflect the results of the query due to the relatively short length of texts processed by the ICD coding system, and the inability to rely solely on TF-IDF to determine word importance. In the ICD coding system, disease codes have their own categories, and the medical descriptions corresponding to codes in the same category will inevitably contain repeated words that are of utmost importance for coding accuracy. However, the similarity calculation based on the vector obtained from the TF-IDF results cannot perfectly match multiple documents with diagnostic text and accurately rank the best matching code. Thus, the bidirectional matching algorithm enhances the initial algorithm by correcting its matching accuracy.

The foundation of the similarity correction algorithm lies in the trust of the Gensim and assumes that the initial similarity values already match the codes with a higher degree of similarity. The correction algorithm recalculates the matching values and reorders the high similarity codes to obtain more accurate similarity rankings, in preparation for the automatic generation of the ICD-11 coding algorithm. The bidirectional matching degree calculation method can be represented by Algorithm 1.

To measure the effectiveness of the matched words using our method, let r represent the number of matched words, m represents the number of words in the diagnosis text, and n represent the number of words in the code text. The bidirectional matching degree v is defined as follows:

$$v = \frac{r^2}{mn} \tag{6}$$

The ultimate corrected similarity score, denoted as η, is defined based on the text similarity model as follows:

$$\eta = u + v \tag{7}$$

where u represents the initial similarity score.

Algorithm 1 Bidirectional Matching Calculation Method Based on TF-IDF Model

Input: Disease text t
Output: A set of codes with similarity estimates

let List_doc_sim ← TFIDF (t)
let List_candidate ← empty list
let List_candidate_sim ← empty list
for each(doc, doc_sim) ∈ list_doc_sim **do**
    if *doc_sim* > 0.2 **then**
        list_candidate.add(*doc*)
        list_candidate.add(*doc_sim*)
    end if
end for
Sort *list_candidate* by *list_candidate_sim* ASC
let *list_t_sim* ← empty list
for each c in *list_candidate* **do**
    list_c ← split(c)
    n_occurrence ← 0
    for each w in split(t) **do**
        if w ∈ list_c **then**
            n_occurrence ++
            t_ sim = n_occurrance / len (list_c) ∗ (n_occurrance /len (split(t))^2
            list_t_sim.add(t_sim)
        end if
    end for
end for
return list_candidate(k) where list_t_sim(k) is maximum

2.4 The bidirectional matching-based similarity correction algorithm

The concept of automatic generation of ICD-11 cluster codes is based on the coding rules of ICD-11 and the results of trust correction similarity values. The most distinctive feature of the ICD-11 coding rules is expansion and coordination. Expansion refers to the two types of ICD-11 extension codes, one is the extension code starting with X, and the other is the extension of the meaning brought by the stem code. The extension code itself does not contain diagnostic information but describes other information about the disease or health status. When one wants to express diagnostic information more simply and flexibly, the ICD-11 diagnostic codes can be combined using the "&" and "/" operators to form a diagnostic sentence. Through an intricate analysis and systematic arrangement of the ICD-11 coding guidelines, we have attained a definitive classification of five distinct coding categories, as shown in Table 1.

Table 1: Code combinations and connection symbols

Category	Code combination	Connection symbol	Example and explanation
1	One stem code	N/A	BA41.0
2	One or more extension codes after one stem code	&	BA41.0 & XA7RE3 & XA7NQ7
3	Two stem codes	/	BA41.0/1A00
4	Complications	/ and &	DD51 & XK8G/ME24.2 & XT5R
5	Two unrelated symptoms	And, /, &	NA07.0/PA60 & XE1DA & XE53A and NC32.2 & XK8G & XJ7YM/PA60 & XE1DA & XE53A

By the coding rules, the algorithm for automatically generating ICD-11codes first sets the disease code with the highest similarity score among valid stem codes as the default value. Then, based on this similarity score, the algorithm sequentially compares it with the extension codes, codes after the first position of valid stem codes, and invalid stem codes, and codes within the defined range are compiled into the final generated code. The specific flowchart of the automatic disease code generation algorithm is illustrated in Algorithm 2.

Algorithm 2 Automated generation of ICD-11 cluster coding

Input: Stem Valid Code S_1, Stem invalid code S_2, Extension Code E
Output: AutoCode of ICD-11

$S_l \leftarrow 0$
AutoCode $\leftarrow$ Code of the first row and column of S_l
if $E = 0$ **then**
  for $i \leftarrow 2$ to n **do**
    if Code of $S_I[i]$ starts with valid code **then**
      Append $S_I[i][1]$ to AutoCode
      return AutoCode
    end if
  end for
else
  for $i \leftarrow 1$ to m **do**
    if $E[i][2] > S_I[0][2]-$ value **then**
      Append "/" and the code of $E[i]$ to AutoCode
    else
      Apply statistical analysis to obtain the value range of $E[i][2]$
      if $S_I[0][2]-$ value $1 \leqslant S_l[i][2] \leqslant S_I[0][2]-$ value 2 **then**
        Append "/" and the code of $S_I[i]$ to AutoCode
      end if
    end if
  end for
  if $S_2 = 0$ **then**
    for $i \leftarrow 1$ to k **do**
      if $S_1[0][2]-$ value $3 < S_2[i][2]$ **then**
        Append "/" and the code of $S_2[i]$ to AutoCode
      end if
    end for
    return AutoCode
  else
    Append "/" and the code of the first row and column of S_2 to AutoCode
    return AutoCode
  end if
end if

3 Results

3.1 The impact of different sample coding on the system performance

In this study, we utilized two distinct sets of test data. The medical record dataset consists of 500 electronic medical records from Chinese hospitals, including the text of disease diagnoses and their recorded ICD-10 codes. It should be noted that the language used in the ICD-10 codes is different from that used in the ICD-11 codes because the latter modify or delete several medical terms. The ICD-11 Browser dataset consists of 100 samples of disease codes based on a combination of master and extended codes extracted from the ICD-11 Browser Website[31], with corresponding Chinese language

description. Table 2 provides examples of the datasets used in our experiments. In addition, to ensure the quality of the ICD-11 coding assessment, three medical experts labeled the diagnostic texts in the dataset with the correct ICD-11 codes. To further assess the reliability among these three experts, we used the Fleiss' kappa value, which is a valid method for assessing reliability[32]. The higher the percentage of overlapping labels on which the assessors agreed, the more reliable the code given. The final inter-assessor agreement (kappa) was determined to be 97.53%.

Table 2: Examples of main items in medical records dataset and the ICD-11 browser dataset

Dataset type	PatientID	Gender	Frequency of medical consultations	Diagnostic Text	Correct Code
medical records	458727	Male	2	Cerebral hemorrhage	8B25.1
medical records	455667	Female	1	Immune thrombocytopenic purpura	3B64.Z
ICD-11 Browser	-	-	-	Multibacillary leprosy	1B20.1
ICD-11 Browser	-	-	-	Other specified acquired immunodeficiencies/Pregnancy	4A20.Y&XT0S

The primary function of the computer-assisted ICD- 11 disease coding system is to utilize the ICD-11 coding library to accurately generate the necessary codes for coders based on the diagnostic text inputted by users. Its performance can be classified as an information retrieval system, and therefore, its performance is reflected in the precision and recall of generating codes for specific tasks. This article utilizes sample data to test and analyze the precision (P), recall (R), and F1 score of the proposed method. After each search, all codes are classified into four groups: true positives (TP) for codes that are correctly retrieved by the system, false positives (FP) for codes that are retrieved but not relevant, false negatives (FN) for codes that are relevant but not retrieved by the system, and true negatives (TN) for codes that are not relevant and not retrieved by the system. Therefore, the calculation methods for P, R, and $F1$ (F1-score) values during the search process are defined as follows:

$$P = \frac{TP}{TP + FP} \tag{8}$$

$$R = \frac{TP}{TP + FN} \tag{9}$$

$$F_1 = \frac{2PR}{P + R} \tag{10}$$

3.2 The impact of different sample coding on the ICD-11 automatic coding system performance

This study aims to evaluate the effectiveness of our proposed ICD-10 automated coding system by analyzing its ICD-11 code results generated for two different sets of test data. Figure 3 shows that during the ICD-11 diagnostic example test, there is a significant difference between the diagnostic text and the disease description stored in the existing ICD-11 coding system. This is mainly due to the system's tendency to output additional candidate codes, such as extended codes based on diagnostic text, which in turn leads to lower accuracy of the method. However, the high recall of 0.97 in the tests on the medical records dataset suggests that this approach achieves a higher level of completeness in the query results, with little possibility of missing potential codes.

In tests on the ICD Browser dataset, each set of diagnostic text was highly matched to text in the system's own codebase, leading to a high level of accuracy in the query results. However, the complexity and length of the descriptions in this sample was higher than in the first test, which had a greater impact on the system's coding recall. The precision of the case dataset test was slightly lower than the ICD Browser dataset test, while the recall was slightly higher. In conclusion, the F1 scores, which reflect the overall test results, show that the proposed method achieves almost identical query results for two representative samples of diagnostic text with different characteristics.

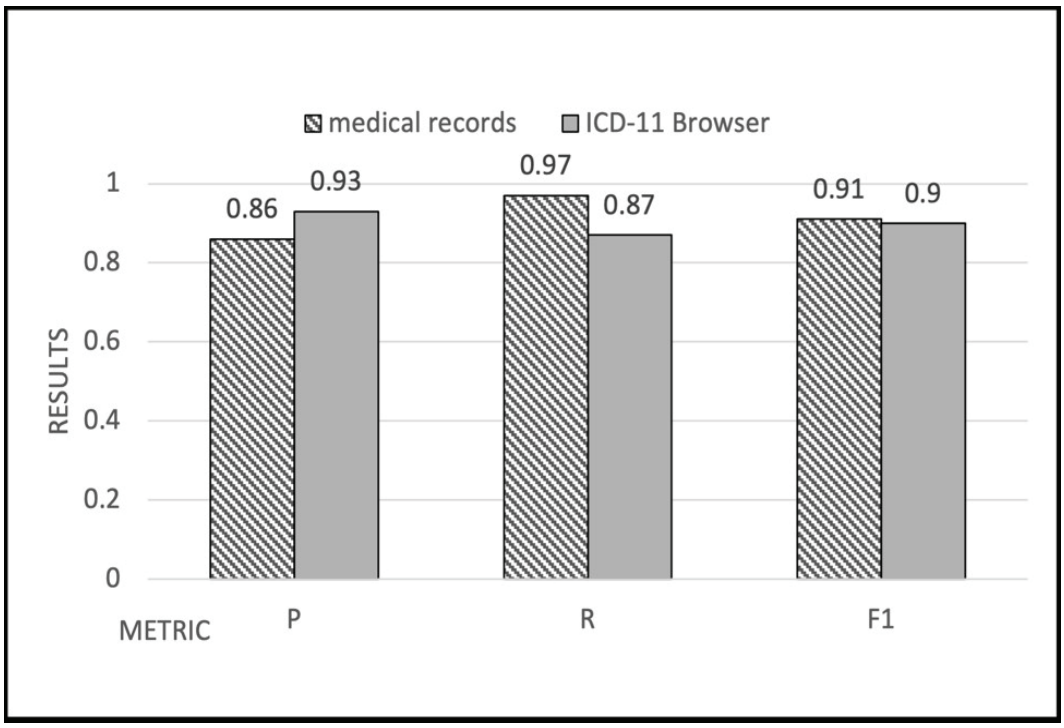


Figure 3: Comparison of results between 80-sample and 50-sample test

3.3 Comparison of the performance between different methods

In the second round of testing, we will use the medical record dataset test to compare the performance of our system with the existing disease coding query system, MedSci. To ensure that the variables are as consistent as possible in the second round of testing, we used a "0" or "1" normalization method to determine the value of the correctly extracted code when calculating the query results returned by the MedSci Disease Coding System.A "0" indicates that the correct code did not appear in the query results, while a "1" indicates that the query results contained the correct code or a similar code. The query results of the MedSci system coding system are shown in Table 3. Figure 4 shows that the query quality of our proposed system is much better than the MedSci system in terms of P, R, and overall F1. The highly matched diagnostic texts in the second round of testing and the system's coding database resulted in a significantly improved precision compared to the first round of testing.

Table 3: Examples of search results from the MedSci system

ID	Diagnostic Text	Correct Code	Results on the MedSci System
1	Malignant esophageal tumor	2B70.Z	2B70, 2B70.Y, 2B70.Z
2	Acute hepatitis B	1E50.1	No results found. Please shorten or change the keywords.
3	Pulmonary emphysema	CA21.Z	CA21, CA21.Y, CA21.Z, CB03.1, KB27.0

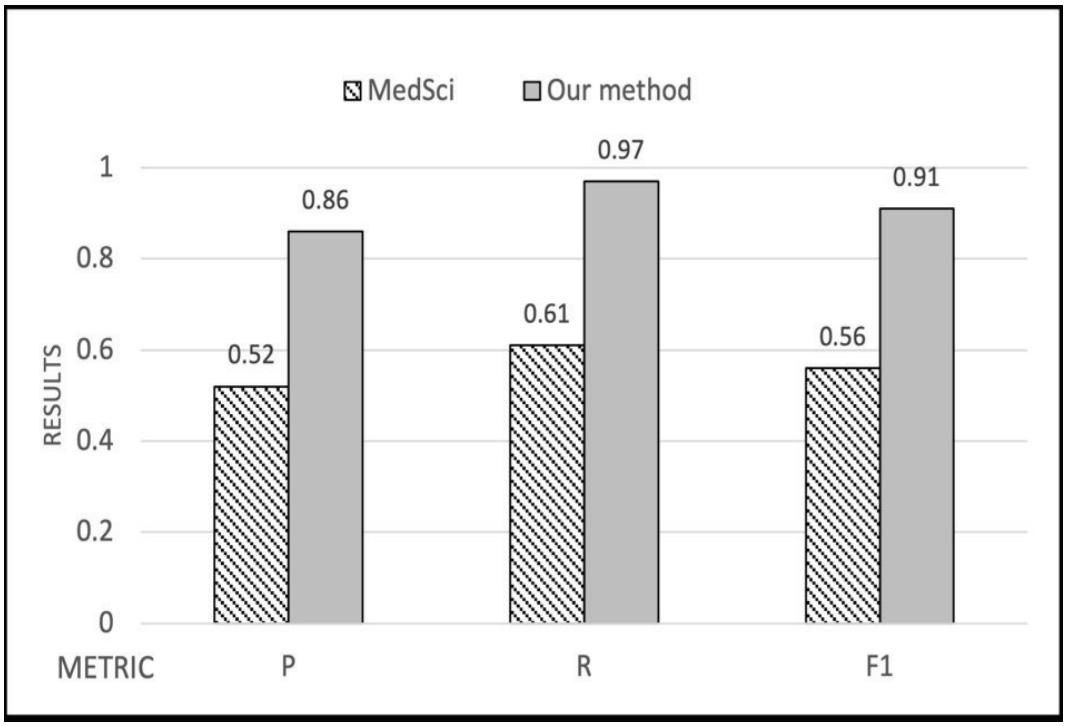


Figure 4: Comparison of results between 80-sample and 50-sample test

3.4 ICD-11 coding system prototype

The innovative method of generating ICD-11 cluster codes based on medical records has achieved better experimental results, and the main interface of our constructed system is shown in Figure 5. The process of using our proposed ICD-11 coding system is as follows. First, the coder inputs the diagnostic information. Next, the coding system provides relevant terms through a semantic dictionary for the coder to complete and refine the diagnostic text. The coder then selects the target code from the list of ICD codes returned. The system provides the coder with supporting information from the ICD-11 chart database related to the target code. In addition, the coding system lists additional ICD codes for the coder to select based on the relevant postpositional relationship of the target code. Based on the selected ICD code or cluster of codes, the system generates cluster codes.

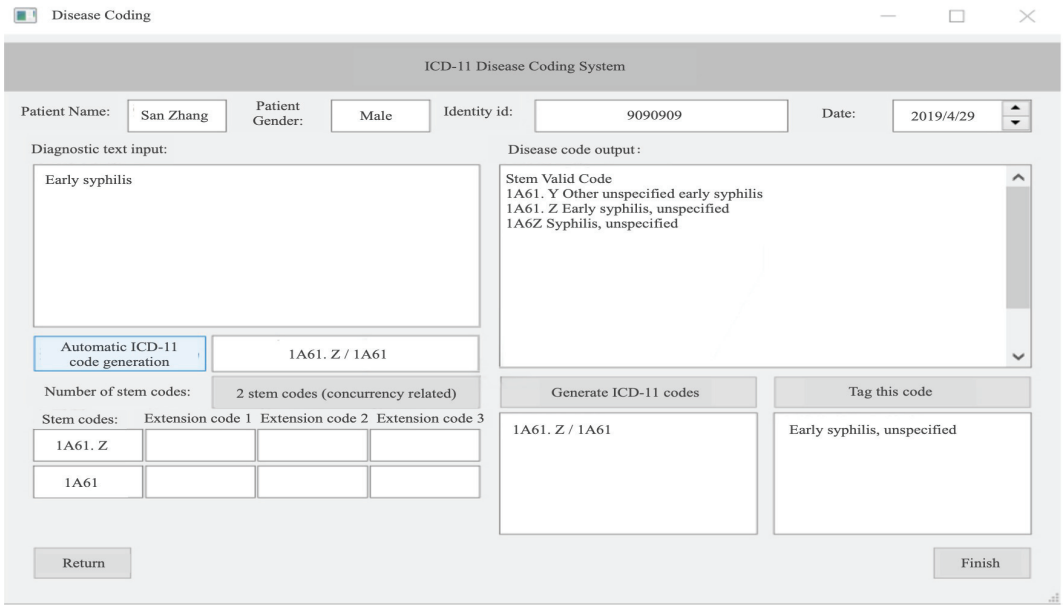


Figure 5: Comparison of results between 80-sample and 50-sample test

4 Discussion

We provide a promising solution for generating ICD-11 cluster codes for automated medical record coding and demonstrate the potential for utilizing a computer-assisted coding system to improve the efficiency and accuracy of ICD-11 coding. To evaluate the performance of our proposed method, we compared the results generated by the ICD-11 automated coding system with the ICD-11 medical record dataset manually annotated by medical experts. The experimental results show that our proposed method has a precision of 0.86 and a recall of 0.97, which is better than existing methods. The method effectively solves the complex cluster coding construction problem in the ICD-11 standard. Notably, the experimental results in Figure 3 demonstrate that the performance of our proposed method is not affected by the features of diagnostic text, and the F1 score can reach more than 0.90, indicating that the method is highly reliable. This suggests that our proposed method is robust and reliable regardless of the complexity and length of the diagnostic text.

In addition, our proposed ICD-11 coding system provides a computer-assisted coding paradigm that demonstrates the use of semantic text similarity computation methods, indexing-based retrieval, and knowledge relation utilization in coding. The system also provides normal and flexible query modes that enhance coding search capabilities. This provides a valuable reference for building effective computer-assisted coding systems. Figure 4 shows that the query quality of the proposed system significantly outperforms the MedSci system in terms of P, R and F1. The system illustrates how the computer-assisted ICD-11 system helps the ICD-11 encoder to generate cluster codes. The performance evaluation metrics introduced by this computer-aided coding method and information retrieval system effectively address the lack of guidance in the early implementation and promotion of the ICD-11 standard.

However, our study still has some limitations. One potential limitation is that ICD-11 is still being tested in a small number of hospitals, and therefore, not enough data have been accumulated for data-driven modeling. ICD-11 provides a more complex coding structure and guidelines for medical coding, and therefore requires a large amount of data to train an accurate model compared to previous versions[33]. As ICD-11 becomes more widely adopted and the ICD-11 corpus expands, the we will use larger datasets for validation, which can effectively improve the performance of the coding system. In addition, the ICD-11 revision is constructed based on the semantic web, and the knowledge relations of the existing version are derived from its core ontology. Therefore, the direct use of its ontology model will greatly contribute to the development of smarter coding systems that utilize semantic knowledge for disease knowledge reasoning, automatic coding of medical records, coding pattern optimization, and self-learning. Nevertheless, the limited information content of the Chinese version of the ICD coding system published by the World Health Organization lacks the necessary model information and knowledge constraints, which limits the ability of the healthcare information industry to build an intelligent computer-aided ICD-11 coding system and to utilize the existing knowledge and proven ontology reasoning engines to complete the coding process.

5 Conclusions

Our study proposes an innovative method for automated ICD-11 coding of medical records. Key technical advancements include bidirectional matching to refine text-code similarity and an algorithm to produce valid ICD-11 cluster codes. Extensive experiments highlight significant improvements in P, R and F1 over baseline methods. The proposed computer-assisted approach can enhance clinical coding efficiency and quality. Limitations exist in testing on a broader range of real-world datasets. Future work can incorporate ICD-11 ontology knowledge and evaluate scaling to larger medical corpora. Overall, this research provides important foundational insights into next-generation clinical informatics using emerging standards like ICD-11. Intelligent coding systems will grow increasingly valuable as healthcare data continues expanding globally.

Funding

This work was funded partially by the National Natural Science Foundation of China with the grant number 62173025, a major project of the National Social Science Foundation of China with the grant number 18ZDA086, and the National Natural Science Foundation of China with the grant number 62102087.

Acknowledgment

The authors appreciate the support of the Beijing Logistics Informatics Research Base.

Conflict of interest

The authors declare no conflict of interest.

References

[1] Biruta Sloka, Anna Angena(2022). Challenges for health care financing in latvia – comparison with other baltic countries, *Journal of Service, Innovation and Sustainable Development*, 3(2), 143–152, 2022.

[2] [Online]. https://icd.who.int/en/docs/icd11factsheet_en.pdf, Accesed on 10 November 2022.

[3] Zhu, V. J., Lenert, L. A., Barth, K. S., Simpson, K. N., Li, H., Kopscik, M., Brady, K. T. (2022). Automatically identifying opioid use disorder in non-cancer patients on chronic opioid therapy, *Health Informatics Journal*, 28(2), 2022.

[4] Eastwood, C. A., Southern, D. A., Doktorchik, C., Khair, S., Cullen, D., Boxill, A., Quan, H. (2021). Training and experience of coding with the World Health Organization's International Classification of Diseases, Eleventh Revision, *Health Information Management Journal*, 52(2), 92–100, 2021.

[5] Saravanan A, Anandhi D, Srividya M. (2023). Class probability distribution based maximum entropy model for classification of datasets with sparse instances, *Computer Science and Information Systems*, 20(3), 949-976, 2023.

[6] Kaur R, Ginige JA, Obst O. (2022). AI-based ICD coding and classification approaches using discharge summaries: A systematic literature review, *Expert Systems with Applications*, 213, 118997, 2022.

[7] Yamada, E., Aramaki, E., Imai, T., Ohe, K. (2010). Internal structure of a disease name and its application for ICD coding, *Studies in Health Technology and Informatics*, 160, 1010-1014, 2010.

[8] Filip, F. G. (2023). Automation and computers and their contribution to human well-being and resilience, *Studies in Informatics and Control*, 30(4), 5-18, 2023.

[9] Nakahara, S., Uchida, Y., Oda, J., Yokota, J. (2014). Bridging classification for injury diagnoses that can be converted to both the International Classification of Diseases and the Abbreviated Injury Scale, *Acute Medicine and Surgery*, 1(1), 10-16, 2014.

[10] [Online]. https://www.medsci.cn/sci/icd-10.do, Accesed on 12 November 2022.

[11] Gill, P. J., Thavam, T., Anwar, M. R., Zhu, J., To, T., Mahant, S. (2022). Pediatric Clinical Classification System for use in Canadian inpatient settings, *Plos one*, 17(8), e0273580, 2022.

[12] Fung, K. W., Xu, J., Ameye, F., Gutiérrez, A. R., Busquets, A. (2018). Re-purposing the ICD-9-CM procedures index for coding in ICD-10-PCS and SNOMED CT, *American Medical Informatics Association Annual Symposium Proceedings*, 2018, 450, 2018.

[13] Venepalli, N. K., Qamruzzaman, Y., Li, J. J., Lussier, Y. A., Boyd, A. D. (2014). Identifying clinically disruptive International Classification of Diseases 10th Revision Clinical Modification conversions to mitigate financial costs using an online tool, *Journal of Oncology Practice*, 10(2), 97-103, 2014.

[14] Ertuğrul D Ç, Abdullah S A. (2022). A Decision-Making Tool for Early Detection of Breast Cancer on Mammographic Images, *Tehnički vjesnik*, 29(5), 1528-1536, 2022.

[15] Hamad, A. F., Vasylkiv, V., Yan, L., Sanusi, R., Ayilara, O., Delaney, J. A., Lix, L. M. (2021). Mapping three versions of the international classification of diseases to categories of chronic conditions, *International Journal of Population Data Science*, 6(1), 1406, 2021.

[16] Cao, L., Gu, D., Ni, Y., **e, G. (2019). Automatic ICD code assignment based on ICD's hierarchy structure for Chinese electronic medical records, *AMIA Summits on Translational Science Proceedings*, 2019, 417-424, 2019.

[17] Fareh, M., Riali, I., Kherbache, H., Guemmouz, M. (2023). Probabilistic reasoning for diagnosis prediction of Coronavirus disease based on probabilistic ontology, *Computer Science and Information Systems*, 20(3), 1109-1132, 2023.

[18] Wu, Y., Chen, Z., Yao, X., Chen, X., Zhou, Z., Xue, J. (2022). JAN: Joint Attention Networks for Automatic ICD Coding, *IEEE Journal of Biomedical and Health Informatics*, 26(10), 5235-5246, 2022.

[19] Teng, F., Zhang, Q., Zhou, X., Hu, J., Li, T. (2024). Few-shot ICD coding with knowledge transfer and evidence representation, *Expert Systems with Applications*, 238, 121861, 2024.

[20] Lee H, Kim S. (2023). Impact of the ICD-11 on the accuracy of clinical coding in Korea, *Health Information Management Journal*, 52(3), 221-228, 2023.

[21] Fung KW, Xu J, Bodenreider O. (2020). The new International Classification of Diseases 11th edition: a comparative analysis with ICD-10 and ICD-10-CM, *Journal of the American Medical Informatics Association*, 27(5), 738-746, 2020.

[22] Eastwood, C. A., Southern, D. A., Khair, S., Doktorchik, C., Cullen, D., Ghali, W. A., Quan, H. (2022). Field testing a new ICD coding system: methods and early experiences with ICD-11 Beta Version 2018, *BMC Research Notes*, 15(1), 1-7, 2022.

[23] Perotte, A., Pivovarov, R., Natarajan, K., Weiskopf, N., Wood, F., Elhadad, N. (2014). Diagnosis code assignment: models and evaluation metrics. *Journal of the American Medical Informatics Association*, 21(2), 231-237, 2014.

[24] Venkatesh R, Shenbagarajan A, Shenbagalakshmi G. (2023). Multi-gradient boosted adaptive SVM-based prediction of heart disease, *International Journal of Computers Communications and Control*, 18(5), 2023.

[25] Wang, Y. (2022). Online Healthcare Privacy Disclosure User Group Profile Modeling Based on Multimodal Fusion, *International Journal of Computers Communications and Control* , 17(5), 2022. Doi: 10.15837/ijccc.2022.5.4696.

[26] Negoiţă, RF, Borangiu T. (2023). Robotic Process Automation of Inventory Demand with Intelligent Reservation, *Studies in Informatics and Control*, 32(2), 5-14, 2023.

[27] Boerma, T., Harrison, J., Jakob, R., Mathers, C., Schmider, A., Weber, S. (2016). Revising the ICD: Explaining the WHO approach, *The Lancet*, 388(10059), 2476-2477, 2016.

[28] Robertson S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF, *Journal of Documentation*, 60(5), 503-520, 2004.

[29] Huang A. (2008). Similarity measures for text document clustering, *Proceedings of the sixth New Zealand computer science research student conference*, 2008, 9-56, 2008.

[30] Gomaa WH, Fahmy AA (2013). A survey of text similarity approaches, *International Journal of Computer Applications*, 68(13), 13-18, 2013.

[31] [Online]. Available: https://icd.who.int/browse11/l-m/en, Accesed on 10 November 2022.

[32] Mousavi, R., Raghu, T. S., Frey, K. (2020). Harnessing artificial intelligence to improve the quality of answers in online question-answering health forums, *Journal of Management Information Systems*, 37(4), 1073-1098, 2020.

[33] Teng, F., Liu, Y., Li, T., Zhang, Y., Li, S., Zhao, Y. (2023). A review on deep neural networks for ICD coding, *IEEE Transactions on Knowledge and Data Engineering*, 35(5), 4357-4375, 2023.

This journal is a member of, and subscribes to the principles of,
the Committee on Publication Ethics (COPE).
https://publicationethics.org/members/international-journal-computers-communications-and-control

Cite this paper as:

Feng, J.; Zhang, R.; Chen, D.; Shi, L.; Li, Z;(2024). Automated Generation of ICD-11 Cluster Codes for Precision Medical Record Classification, *International Journal of Computers Communications & Control*, 19(1), 6251, 2024.
https://doi.org/10.15837/ijccc.2024.1.6251