

Construction of a Semi-automatic ICD-10 Coding System

Lingling Zhou

Department of Information, Daping Hospital of Army Medical University

Cheng Cheng

Department of Information, Daping Hospital of Army Medical University

Dong Ou

Department of Information, Daping Hospital of Army Medical University

Hao Huang (✉ m13608388426@163.com)

Army Medical University Daping Hospital and Research Institute of Surgery, Department of Information

<https://orcid.org/0000-0003-1315-7077>

Research article

Keywords: ICD-10 coding – Automatic coding-Description models of the regular expressions - Diagnosis codes - Diagnosis descriptions

Posted Date: April 1st, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-19839/v2>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Version of Record: A version of this preprint was published on April 15th, 2020. See the published version at <https://doi.org/10.1186/s12911-020-1085-4>.

Abstract

Background

The International Classification of Diseases, 10th Revision (ICD-10) has been widely used to describe the diagnosis information of patients. Automatic ICD-10 coding is important because manually assigning codes is expensive, time consuming and error prone. Although numerous approaches have been developed to explore automatic coding, few of them have been applied in practice. Our aim is to construct a practical, automatic ICD-10 coding machine to improve coding efficiency and quality in daily work.

Methods

In this study, we propose the use of regular expressions (regexps) to establish a correspondence between diagnosis codes and diagnosis descriptions in outpatient settings and at admission and discharge. The description models of the regexps were embedded in our upgraded coding system, which queries a diagnosis description and assigns a unique diagnosis code. Like most studies, the precision (P), recall (R), F-measure (F) and overall accuracy (A) were used to evaluate the system performance. Our study had two stages. The datasets were obtained from the diagnosis information on the homepage of the discharge medical record. The testing sets were from October 1, 2017 to April 30, 2018 and from July 1, 2018 to January 31, 2019.

Results

The values of P were 89.27% and 88.38% in the first testing phase and the second testing phase, respectively, which demonstrate high precision. The automatic ICD-10 coding system completed more than 160,000 codes in 16 months, which reduced the workload of the coders. In addition, a comparison between the amount of time needed for manual coding and automatic coding indicated the effectiveness of the system-the time needed for automatic coding takes nearly 100 times less than manual coding.

Conclusions

Our automatic coding system is well suited for the coding task. Further studies are warranted to perfect the description models of the regexps and to develop synthetic approaches to improve system performance.

Background

The [International Statistical Classification of Diseases and Related Health Problems \(ICD\)](#)[1] is the foundation for the identification of health trends and statistics globally and the international standard for reporting diseases and health conditions. The ICD has been published in a series of editions to reflect advances in health and medical science over time. The 10th version, ICD-10, was endorsed in May 1990 and is used by more than 100 countries around the world. In China, ICD-10 was adopted in 2001 by the Peking Union Medical College Hospital of the World Health Organization (WHO) Collaborating Centre for the Family of International Classifications. The Chinese Version of Classification and Codes of Diseases (CCD) (GB/T 14396-2001), is an expanded version of ICD-10 and is widely used in China. In 2011, the National Health and Family Planning Commission of the People's Republic of China promulgated the notice on using CCD codes for encoding diagnoses in the medical records after January 1, 2012[2]. However, the vast area and large population of China

have caused increases in numerous local diseases and special diseases, resulting in the localized ICD-10 coding systems have been widely explored by health departments of various provinces and cities. As a military hospital, we adopt the uniform coding rules based on the ICD-10 issued by the military management department, and add some additional codes according to the actual situation of our hospital. Although a version of the ICD-11 code set was released on June 18, 2018 to allow Member States to prepare for its implementation by the WHO, our military unit is still committed to using the ICD-10 to lay the foundation for using ICD-11.

Diagnosis-related group (DRG) is defined as classifications of each patient according to the diagnosis and other characteristics of the case. Diagnosis-related group-based prospective payment system (DRG-PPS) is one type of advanced hospital payment reimbursement mechanism. Since the 1990s, payments based on DRG have gradually become the principal means of reimbursing acute inpatient care in most high-income countries[3, 4]. Approximately a decade later, more and more low- and middle-income countries have begun to establish DRG-based payment systems. In July 2011, the pilot project of DRG was launched in 6 hospitals in Beijing, China. On June 2, 2017, the National Health Commission of the People's Republic of China held a pilot meeting on DRG-PPS reform in Shenzhen, Guangdong Province. The reform of payment systems in public hospitals has entered a crucial stage in our country. The most commonly used diagnosis classification standard in DRG is the ICD-10 code system, and the quality of coding based on ICD-10 codes is directly related to the DRG grouping, which directly affects the economic benefits of the hospital.

Manually assigning ICD codes is expensive, time consuming and error prone due to the vast coverage and extent of ICD terminology. Many scientists have explored automatic and semi-automatic ICD coding as a solution to the massive amounts of human labour required for manual coding. For example, several studies based on machine learning approaches, such as the support vector machine (SVM) method[5-8], were proposed to automatically assign ICD-10 codes. With the extensive application of deep learning methods in various fields, these methods have also been widely used in automated ICD coding [9-12]. These studies indicate that deep learning models can produce interpretable results and can code automatically in a reasonable way. Meanwhile, studies [9, 10] also indicate that the deep learning framework outperforms SVMs. Other techniques based on natural language processing (NLP)[13-17] can significantly increase the performance of automatic ICD coding by mapping the already assigned diagnoses of patients to ICD codes. Study [18] used word embedding combined with a convolutional neural network (CCN), which showed outstanding performance compared with the NLP plus supervised machine learning models. A prior study [19] automatically classify patients' diseases into an ICD-10-CM category based on the well known Web Ontology Language. Another study[20] combined three coding systems into a single superior system to improve the performance of automated ICD-9-CM codes in clinical reports. In [12, 21-22], ICD-10 coding of death certificates has already been addressed by automation. In [23] semi-automatic assignment of the indexing system was performed by exploiting the idea of the bag-of-words mapping indexing tool. Most of the above methods are only theoretical, and they have not yet been applied in practice.

At present, there are mainly two ways of coding in our country. One way is that clinicians are required to learn the ICD-10 codes and enter the clinical diagnosis with the help of a navigation tool within the nomenclature according to the coding principle. Clinicians completing patient records can use this tool to obtain a preliminary coding of the record, and then the medical record coders perform the quality audit. The other way of coding is

that the medical record coders code according to the clinician's diagnosis records by using the computer dictionary library and then communicate with the clinicians when disputes arise. The first method has a higher coding efficiency, and the other method has a higher coding quality. The first method increases the workload of clinicians and requires better information technology for hospitals, so ICD-10 coding in many hospitals is still undertaken by the coders in medical records departments according to the free text form in the clinician's diagnosis records. Our hospital also adopts the second method. There are several types of errors in this method. First, clinicians often utilize abbreviations or synonyms, which causes imprecision and ambiguity when the coders are matching ICD-10 codes to these diagnosis descriptions. Second, several diagnosis descriptions are closely related and should be combined in many cases into a single combination ICD-10 code. However, coders may code each disease separately. Third, ICD-10 codes are organized in a hierarchical structure where the top-level codes represent generic disease categories and the bottom-level codes represent more specific diseases [24], and the coder may match the diagnosis description to a generic code instead of a specific code. In addition, the coders could make obvious errors due to carelessness when the workload is heavy.

There is a limited set of diseases that hospitals can treat. Different clinicians have their own recording habits. Through statistical analysis, their regularity can be determined. The ICD coding set has clear coding rules that are suitable for clustering and matching by related technologies. To improve coding efficiency and quality and to reduce coding errors, we aim to build an automatic ICD-10 coding machine on the basis of the existing coding system, which translates the free-text diagnosis descriptions into ICD-10 codes.

Methods

Coding System

Our hospital, as a member of a large-scale comprehensive medical institution in Chongqing, China, is involved in medical care, preventive health care, education and scientific research. The annual average number of discharged patients is approximately 100,000. A large number of discharge medical records are waiting to be coded by coders. The coding system-we currently use is called the Medical Record Cataloging System—a subsystem of the hospital information system. It consists of general sociodemographic information, diagnosis information, surgical information, and cost information. All of the above information forms the homepage of the discharge medical records. The general sociodemographic information is filled in by the staff at the admissions office when the patient is admitted, while the cost information is completed by the staff of the cost office when the patient is discharged from the hospital. Diagnosis information includes the diagnosis descriptions of outpatients, admissions, discharges, and their corresponding diagnosis codes, treatment results, and treatment days. Operation information includes the operation descriptions, operation codes, surgeons and anaesthesiologists, and operating date. After clinicians record these descriptions, the coders complete the corresponding codes with the aid of the ICD-10 dictionary library embedded in the Medical Record Cataloging System. In this study, only diagnosis descriptions and corresponding diagnosis codes were used.

Constructing the Upgraded Coding System

The description models of the regular expressions (regexps) were applied to automatically transform the diagnosis descriptions to the matching codes in the upgraded coding system. The following steps specifically describe how to build the system.

Step 1: Data pre-processing. The diagnosis descriptions of the modeling datasets are classified based on the diagnosis codes. When the same diagnosis description corresponds to multiple diagnosis codes or multiple diagnosis descriptions correspond to one diagnosis code, the one that most coders agree on is used. According to the ICD-10 coding guidelines, each diagnosis code matches a diagnosis description.

Step 2: Diagnosis terminology parsing. The nonstandard diagnosis descriptions representing the same disease recorded by clinicians were transformed to a diagnosis description through the description models of the regexps. A description model of the regexps matches an ICD-10 code. The regexps [25] replace the usual percentage (%) and other similarity lookup methods by using some pre-agreed combination of regular special symbols, such as ^, *, ., x|y, or ?, as shown in Table 1. Each special symbol represents a different meaning and is combined into the regexps. For example, complete abortion, foetal malformation, abdominal pain, haemorrhagic anaemia, ect, were translated into regexps, and we show these in Chinese due to the diversity of diagnosis descriptions recorded by clinicians in our country (Figure 1). All the description models of the regexps were generated manually by our information engineers in cooperation with the coders. A rule base consists of these established description models of the regexps is used for subsequent program running.

Step 3: Automatic coding. Diagnosis descriptions waiting for coding after the clinician fills in the homepage are automatically stored in Oracle Database 10g Software[26]. The REGEXP_LIKE function supported by Oracle[27] is used to complete the one-to-many matching between the transformed diagnosis descriptions and an ICD-10 code. If the coding matching is completely consistent, the diagnosis code is automatically generated, and the coding log is updated. If the diagnosis description cannot be matched or corresponds to multiple regexps, the coding failure log is generated.

Step 4: Code auditing. The coders check the coincidence of the automatically completed diagnosis codes one by one through the Medical Record Cataloging System and process the failed codes, including repeated codes, mismatch codes and loss codes. These errors are fed back to the programmer to modify the regexps. Eventually, the regexps are gradually optimized to reduce system coding errors.

The flow chart of automatic coding is shown in Figure 2. Procedural codes and structured query language (SQL) statements are shown in Additional file 1.

Datasets

Our datasets were obtained from the diagnosis information on the homepage of the discharge medical records. The study consisted of two stages, which were divided according to the coding date and time. In the first stage, the top 500 high-frequency diagnosis code categories from January 1, 2014 to June 30, 2017 were used to establish the description models of the regexps. Modeling and optimization lasted three months. The test performance between October 1, 2017 and April 30, 2018 proved that the first phase of the study was effective. Therefore, further research was carried out to increase the total number of matching code categories (up to 1000) in the second stage. The other unmatched (in the first stage) top 500 high-frequency codes between January 1, 2014 and April 30, 2018 were modeled. The second phase of the experiment lasted two months. Testing results from July 1, 2018 to January 31, 2019 were used to further suggest the effectiveness of this study.

Evaluation Metrics

System performance[20, 28] was evaluated using precision (P), recall (R), the F-measure (F) and overall accuracy (A). P is defined as the ratio of true positive (TP) to the total number of TP and false positive (FP). R is the ratio of TP to the total number of TP and false negative (FN). F represents the harmonic mean of the precision and recall, where α is set to 0.5, indicating that equal weight is given to the precision and recall. A is the fraction of coding results assigned correctly among all the codes (TN represents true negative). The definition and equations are shown in Table 2 and Equation 1-4.

$$P = \frac{TP}{TP+FP} \quad (1)$$

$$R = \frac{TP}{TP+FN} \quad (2)$$

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha) \frac{1}{R}} \quad (3)$$

$$P = \frac{TP}{TP+FP} \quad (4)$$

Results

The experimental results of the first stage are as follows. From January 1, 2014 to June 30, 2017, the total number of patients discharged was 383,440, with a total of 2,380,623 diagnosis codes and 8,074 code categories (see Additional file 2). An average of 6 diagnosis descriptions per patient was coded by coders. Figure 3 shows the distribution of the top 500 diagnosis codes with high frequency, which indicates that the high-frequency codes are concentrated in the top 100. The top three codes were I25.101, I10.06 and E11.901, which appeared 87,008, 49,128 and 44,430 times and represent coronary atherosclerotic heart disease, essential hypertension grade III and type 2 diabetes mellitus, respectively. According to the ICD-10 classification, Figure 4 shows the histogram of the number of code categories per cluster, which shows that class K contains the most code categories and class P contains the least. Figure 5 shows the histogram of the number of diagnosis codes per cluster, and class I contains the most codes (387,996). The evaluation results of the first test phase are presented in Table 3, which shows high precision.

Figure 6 shows the monthly trends in coding quantity, and curves A and B represent the correctly assigned diagnosis codes by the code auditors and the automatically completed codes by the coding system, respectively, in every month of testing phase. The two curves are very similar, which indicates high precision. A comparison between the time needed for manual coding and automatic coding is shown in Figure 7.

In the second stage, the other unmatched (in the first stage) top 500 high-frequency codes from January 1, 2014 to April 30, 2018 are presented in Figure 8 and in the Additional file 2. In fact, only 950 code categories could be generated automatically in our whole experiment, and the specific reasons will be described in the Discussion section. Table 4 shows the code categories that cannot be described by regexps. The monthly

trends in coding quantity and a comparison between the amount of time needed for coders and the automatic coding system are presented in Figure 6 and Figure 7, respectively. Automatic coding in the second stage still has high precision and efficiency (Table 3).

Discussion

To our knowledge, this study was the first to develop and apply regexps in automatic coding, with the specific purpose of improving coding quality and efficiency. We constructed the description models of the regexps and inserted them into the coding system via the Oracle software. The automatic ICD-10 coding system completed more than 160,000 codes in 16 months, which reduced the workload of coders and showed high precision and efficiency.

Figure 3 indicates that the code categories are concentrated in the top 100 and that perfecting the corresponding description models of the regexps can reduce the number of FNs to improve the R values. Figure 8 shows that the difference of the frequency and variation range between the codes is not as large as Figure 3, which is the main reason we only study the top 1000. Figure 4 and Figure 5 show that diseases of the digestive system and circulatory system in our hospital are the most diverse and largest in number, which indicates that these two kinds of diseases need more attention in the process of building the description models of the regexps. In addition, class Z diseases are the second largest because our hospital has a large neoplasm treatment centre, involving many special screening examinations (Z12), follow-up examinations after treatment (Z08) and radiotherapy and chemotherapy sessions (Z51) for neoplasms. The curve A in Figure 6 represents the diagnosis codes correctly assigned by the automatic coding system in every month of the two testing stages. Despite the downward trend in the first testing stage, in every month, the distances between the curve A and B remained stable; that is, the number of TPs were stable. The quantity of automatic coding decreased because of changes in hospital management, resulting in the system not running for some days. Figure 7 shows the time needed for automatic coding takes nearly 100 times less than manual coding, which clearly presents automatic coding can save much time.

The values of P for the first and second test stages were up to 89.27% and 88.38%, respectively. However, two main factors result in low R, F, and A values. First, automatic coding can only be executed when the programmer starts the program. Currently, it can only be run twice a day: starting working in the morning (8:00 am) and in the afternoon (14:30). Because clinicians usually complete the homepage of the discharge medical records at the end of their work, the number of diagnosis descriptions waiting to be coded peaks at these two times. Starting the program at these times can realize the value of automatic coding very well. At the same time, the coders are also manually coding. When the program stops, these diagnosis descriptions that should be automatically coded are actually completed by the coders. This leads to too many FNs. The more FNs, the smaller the R value is, and the smaller the R value is, the smaller the F value. Second, of more than 8000 code categories in our hospital, we only matched 950 code categories with high frequency, that is about 7000 code categories with frequencies below 300 have been lost. Table 3 shows the unmodeled code categories produced about 300,000 missing codes in 16 months from 10/1/2017 to 1/31/2019, which made the number of TNs large. The high negative values correspond to the low positive values; that is, the accurately assigned codes are few, and the A values are relatively low. Nevertheless, the values of R, F and A increased in the second testing stage, which illustrated that expanding the total number of matching codes was effective. Table 4 shows that

the corresponding description models of the regexps failed to establish 50 code categories (on the top 1000), which were mainly concentrated in factors influencing health status and contact with health services (class Z) and neoplasms (classes C and D). In addition, other code categories are unspecified. The main reason for this result is that diagnosis descriptions recorded by clinicians are not standardized and vary greatly for these diseases, so the correct diagnosis cannot be coded until the coders consult the complete electronic medical record. The results suggest that clinicians need to strengthen their standardization of diagnosis descriptions when recording diagnoses, especially for classes C, D and Z diseases, while programmers and coders should spend more time on these diseases when building models. On the whole, our system has high precision. With the participation of programmers, clinicians and coders, the accuracy of the system can be improved by focusing on the high-frequency diseases and code categories and repeatedly improving the quality and quantity of regexps.

In recent years, although many studies have focused on automatic ICD coding, we want to highlight the following advantages presented by our study. First, compared to other theoretical studies on model validation using public databases[29-31], we use our hospital data for research to make a system that can be directly applied to practical work. Second, coders could identify their own shortcomings and strengthen communication with clinicians in the audit process to improve their accuracy. Third, our hospital receives a large number of doctors for standardized trainings and refreshers every year. Our doctors record diagnosis descriptions in a variety of ways, so our description models of the regexps have strong representativeness and applicability. Fourth, the regexps represent rules that can be easily understood by workers, which requires less involvement of experts in system implementation and can improve the applicability to small-scale medical institutions with more limited information technology. Five, we update the existing manual coding system based on the rule base of regexps to reduce workload and improve the work quality of coders. The technical requirements and computational cost are less than those of the other methods found in most studies[7],[11],[32-36]. CNN [18],[34-36] is one of the state of the art proposals to solve the problem of automatic ICD coding. Despite their high accuracy, there is still a long way to go before they can be used in practice. Our automatic coding system has been running steadily, which can solve the main problems faced by most medical institutions at present - large amount and repetitive coding. Our system is designed and completed in a relatively short time by our own programmers, which runs in a simple environment. Unlike the complex methods described above, they often require the assistance of engineers of information company. The description models of regexps, we have established have good representativeness and can be used for reference. Overall, our method can transfer to other institutions. Programmers can modify these regexps slightly according to actual situation and write them into existing coding system to run.

There are also shortcomings in our study. First, the automatic coding program runs twice a day: once in the morning and once in the afternoon. When the program is not running, coders are required to do manually input the codes. The next step of our study is to explore how to automatically code the diagnosis immediately after the clinician completes the records. Second, coders are required to perform the last step of auditing, so only semi-automation can be achieved. Code auditing puts forward higher requirements for the ability of coders, and coders should continue to participate in relevant professional training and learning. Standardized diagnosis descriptions are beneficial to improve the correctness of coding. The ICD-10 classification data of some error-prone codes can be sent to the relevant clinical departments, which arouses the attention of clinicians to the standardized writing of discharge diagnosis descriptions. Whether a gold standard can be established for

automatic coding auditing remains to be studied. Third, it is hard to build the description models of the regexps for identical diseases with too different diagnosis terms. Our study is based on the diagnosis of common diseases (the top 1000) and fails to include uncommon diseases. Therefore, in future work, with the complete ICD-10 coding set as the goal, matching rules need to be improved constantly. In addition, the recall, F-measure and accuracy are low in our study compared to these methods mentioned above [34-36]. For example, the CNN based method had reached a F-measure of 60.86% with high efficiency [34], and the reference [36] building a feature matrix, by a pretrained word embedding model used to train a CNN had a high testing accuracy (F-measure 90.86%). Whether our system can be fully automated with high precision by combining with the state of the art is a long-term task that we need to consider.

Conclusion

The description models of the regexps can be used to match diagnosis descriptions and ICD codes, which are well-suited for the coding task. The proposed upgraded coding system is feasible and practical for the automatic coding of ICD-10 codes. Further studies are warranted to perfect the description models of the regexps and to develop synthetic approaches to improve system performance.

Abbreviations

ICD: International Statistical Classification of Diseases and Related Health Problems; CCD: Chinese Version of Classification and Codes of Diseases; DRG: Diagnosis-related group; DRG-PPS: Diagnosis-related group-based prospective payment system; SVM: support vector machine; CCN: convolutional neural network; NLP: natural language processing; CNN: convolutional neural network; regexps: regular expressions; TP: true positive; FP: false positive; FN: false negative; TN: true negative

Declarations

Ethics approval and consent to participate

Ethical approval was given by the medical ethics committee of Daping Hospital of Army Medical University with the following reference number: ME (2014), No. fifty-one. Informed consent was waived because this research did not involve individual data.

Consent to publish

Not applicable

Availability of data and materials

All data generated or analyzed during this study are included in the Additional file 2.

Competing interests

The authors declare that they have no competing interests.

Funding

No funding was received for this study.

Authors' contributions

HH and LLZ conceived the study. LLZ and CC wrote and edited the manuscript. LLZ and DO participated in data collection and analysis. LLZ, HH and CC participated in interpretation of results. All authors read and approved the final manuscript.

Acknowledgments

The authors would like to thank all of coders and R&D team members from Information Department and doctors from Clinical departments, for their collecting, providing and analyzing data.

References

1. World Health Organization. ICD purpose and uses. <http://www.who.int/classifications/icd/en/>. 2018. Accessed 10 Dec 2018.
2. Jia Z, Qin W, Duan H, Lv X, and Li H. A Hybrid Method for ICD-10 Auto-Coding of Chinese Diagnoses. *Stud Health Technol Inform*. 2017;245:427-31.
3. Mathauer I, and Wittenbecher F. Hospital payment systems based on diagnosis-related groups: experiences in low- and middle-income countries. *Bull World Health Organ*. 2013;91(10):746-56.
4. Mathauer I, and Wittenbecher F. DRG-based payment systems in low-and middle-income countries: Implementation experiences and challenges. https://www.who.int/health_financing/documents/dp_01_2012_drg_based_payment_systems_in_lmic.pdf?crazycache=1. 2012. Accessed 18 Oct 2018.
5. Wang S, Li X, Chang X, Yao L, Sheng Q, and Long G.. Learning Multiple Diagnosis Codes for ICU Patients with Local Disease Correlation Mining. *ACM Trans Knowl Discov Data*. 2017;11(3):1-21
6. Perotte A, Pivovarov R, Natarajan K., Weiskopf N, Wood F. and Elhadad N. Diagnosis code assignment: models and evaluation metrics. *Jam Med Inform Assn*. 2014;21(2):231-7.
7. Ferrão J, Janela F, Oliveira M, and Martins H. Using structured EHR data and SVM to support ICD-9-CM coding. *IEEE International Conference on Healthcare Informatics*. 2013; doi: [10.1109/ICHI.2013.79](https://doi.org/10.1109/ICHI.2013.79).
8. Atutxa A, Perez A, Casillas A, Atutxa A, Perez A, and Casillas A. Machine Learning Approaches on Diagnostic Term Encoding With the ICD for Clinical Documentation. *IEEE J Biomed Health*. 2018;22(4):1323-29.
9. Zeng M, Li M, Fei Z. Yu Y, Pan Y, and Wang J. Automatic ICD-9 coding via deep transfer learning. *Neurocomputing*. 2019;324(9):43-50.
10. Yu Y, Li M, Liu L, Fei Z, Wu F, and Wang J. Automatic ICD code Assignment of Chinese Clinical Notes based on Multilayer Attention BiRNN. *J Biomed Inform*. 2019; 91:103114.
11. Duarte F, Martins B, Pinto C, and Silva M. A Deep Learning Method for ICD-10 Coding of Free-Text Death Certificates. *Progress in Artificial Intelligence*. 2017;137-49.
12. Duarte F, Martins B, Pinto C, and Silva. Deep neural models for ICD-10 coding of death certificates and autopsy reports in free-text. *J Biomed Inform*. 2018;80:64-77.

13. Chen Y, Lu H, and Li L. Automatic ICD-10 coding algorithm using an improved longest common subsequence based on semantic similarity. *PLoS one*. 2017;12(3):e0173410.
14. Pereira L, Rijo R, Silva C, and Agostinho M. ICD9-based Text Mining Approach to Children Epilepsy Classification. *Procedia Technology*. 2013;9:1351-60.
15. Zhong Q, Karlson E, Gelaye B, Finan S, Avillach P, Smoller J, Cai T, and Williams M. Screening pregnant women for suicidal behavior in electronic medical records: diagnostic codes vs. clinical notes processed by natural language processing. *BMC Med Inform Decis*. 2018;18(1):30.
16. Nguyen A, Truran D, Kemp M, Koopman B, Conlan D, O'Dwyer J, Zhang M, Karimi S, Hassanzadeh H, Lawley M et al. Computer-Assisted Diagnostic Coding: Effectiveness of an NLP-based approach using SNOMED CT to ICD-10 mappings. *AMIA Annu Symp Proc*. 2018:807-16.
17. Kaur R, and Ginige J. Comparative Analysis of Algorithmic Approaches for Auto-Coding with ICD-10-AM and ACHI. *Stud Health Tech Informat*. 2018;252:73-9.
18. Lin C, Hsu C, Lou Y, Yeh S, Lee C, Su S and Chen H. Artificial Intelligence Learning Semantics via External Resources for Classifying Diagnosis Codes in Discharge Notes. *J Med Internet Res*. 2017;19(11): e380.
19. Roldán-García M, García-Godoy M, and Aldana-Montes J. Dione: An OWL representation of ICD-10-CM for classifying patients' diseases. *J Biomed Semant*. 2016;7:62.
20. Martins R, Buchalla C. Automatic coding and selection of causes of death: an adaptation of Iris software for using in Brazil. *Rev Bras Epidemiol*. 2015;18(4):883-93.
21. Mujtaba G, Shuib L, Raj R, Rajandram R, Shaikh K, and Al-Garadi M. Automatic ICD-10 multi-class classification of cause of death from plaintext autopsy reports through expert-driven feature selection. *PLoS one*. 2017;12(2): e0170242.
22. Almagro M, Martínez R, Montalvo S, and Fresno V. A Cross-lingual Approach to Automatic ICD-10 Coding of Death Certificates by Exploring Machine Translation. *J Biomed Inform*. 2019;94:103207.
23. Arifoğlu D, Deniz O, Aleçakır K, and Yöndem M. CodeMagic: Semi-Automatic Assignment of ICD-10-AM Codes to Patient Records. *Information Sciences and Systems*. 2014; 259-68.
24. Ning W, Yu M, and Zhang R. A hierarchical method to automatically encode Chinese diagnoses through semantic similarity estimation. *BMC Med Inform Dec*. 2016;16:30.
25. Xu C, Su J, and Chen S. Exploring efficient grouping algorithms in regular expression matching. *PLoS one*. 2018;13(10): e0206068.
26. Jiang Y. Application and implementation of regular expression in Oracle. *Inform Tech*. 2012;36(1): 1869-71.
27. Stephen S, Chen J, Davidson M, Thomas S, and Trute B. Oracle Database 10g: a platform for BLAST search and Regular Expression pattern matching in life sciences. *Nucleic Acids Res*. 2005;33: D675-9.
28. Sokolova M, and Lapalme G.. A systematic analysis of performance measures for classification tasks. *Infor Process Manas*. 2009; 45(4):427-37.
29. Hou J, Tan M, Stidham R, Colozzi J, Adams D, El-Serag H, and Waljee A. Accuracy of diagnostic codes for identifying patients with ulcerative colitis and Crohn's disease in the Veterans Affairs Health Care System. *Digest Dis Sc*. 2014;59(10) :2406-10.

30. Kavuluru R, Rios A, and Lu Y. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. *Artif Intell Med.* 2015; 65(2):155-66.
31. Adler P, Rimma P, Karthik N, Nicole W, Frank Wand Noémie E. Diagnosis code assignment: models and evaluation metrics. *J Am Med Inform Assoc.* 2014; 21(2): 231–37.
32. Sollie A, Sijmons R, Lindhout D, Ploeg A, Rubio M, Smit G., Verheijen F, Waterham H, Weely D, Wijburg F, et al. A new coding system for metabolic disorders demonstrates gaps in the international disease classifications ICD-10 and SNOMED-CT, which can be barriers to genotype-phenotype data sharing. *Human mutation.* 2013; 34(7): 967-73.
33. Berndorfer S, and Henriksson A. Automated Diagnosis Coding with Combined Text Representations. *Stud Health Tech Informat.* 2017; 235:201-5.
34. Deng Y, Sander A, Faulstich Land Denecke K. Towards automatic encoding of medical procedures using convolutional neural networks and autoencoders. *Artif Intell Med.* 2019;93:29-42.
35. Rios A, and Kavuluru R. Neural transfer learning for assigning diagnosis codes to EMRs. *Artif Intell Med.* 2019;96:116-22.
36. Cao L, Gu D, Ni Y, and Xie G.. Automatic ICD Code Assignment based on ICD's Hierarchy Structure for Chinese Electronic Medical Records . *AMIA Jt Summits Transl Sci Proc.* 2019;2019:417-24.

Tables

Table 1. Examples of Regular Special Symbols and Meaning

Regular symbols	Meaning
^	Start position of string
*	The front character or expression 0 or more times.
.	Any single character other than null
x y	X or Y, where x and y are one or more characters
?	Match the previous sub-expression zero or once

Table 2. The Definition of Evaluation Result

Display ICD-10	Condition positive	Condition negative
Automatically display	TP (Correct result)	FP (Unexpected result)
Doesn't automatically display	FN (Missing result)	TN (Correct absence of result)

Table 3. Evaluation Results of Two stage

age	Testing sets	TP	FP	FN	TN	P(%)	R(%)	F(%)	A(%)
rst	10/1/2017-4/30/2018	50084	6022	161478	67217	89.27	23.67	37.42	41.19
cond	7/1/2018-1/31/2019	85838	11291	222938	82994	88.38	27.90	42.41	41.89

Table 4. Codes of Uncorresponding the Description Models of the Regexps in the ICD-10

Codes	Diagnosis descriptions	Codes	Diagnosis descriptions
C13.901	Malignant neoplasm of hypopharynx, unspecified	R90.001	Intracranial space-occupying lesion
C25.001	Malignant neoplasm of head of pancreas	S00.803	Superficial injury of face
C71.101	Malignant neoplasm of frontal lobe	S01.806	Open wound of face
C77.004	Secondary malignant neoplasm of supraclavicular lymph nodes	Z08.701	Follow-up examination after combined treatment for malignant neoplasm
C79.806	Secondary malignant neoplasm of neck	Z08.101	Follow-up examination after radiotherapy for malignant neoplasm
C79.826	Secondary malignant neoplasm of pelvis	Z08.202	Follow-up examination after chemotherapy for malignant neoplasm
C79.838	Secondary malignant neoplasm of blood vessels	Z08.001	Follow-up examination after surgery for malignant neoplasm
C83.302	Diffuse non-Hodgkin's lymphoma of large cell	Z09.001	Follow-up examination after surgery for other conditions
D12.601	Benign neoplasm of colon, unspecified	Z47.002	Removal of internal fixation device for fracture
D12.801	Benign neoplasm of rectum	Z48.901	Surgical follow-up care, unspecified
D18.001	Intracranial hemangioma	Z51.005	Radiotherapy for malignant neoplasm of oesophagus
D18.037	Hemangioma of limbs	Z51.008	Radiotherapy for neoplasm of brain
D32.008	Benign neoplasm of cerebral meninges of frontal lobe	Z51.013	Radiotherapy for malignant neoplasm of cervix
D33.305	Benign neoplasm of auditory nerve	Z51.109	Chemotherapy for neoplasm of brain
D48.003	Neoplasm of bone (uncertain or unknown behaviour)	Z51.118	Chemotherapy for malignant neoplasm of pancrea
D73.002	Asplenia, postsurgical	Z51.125	Chemotherapy for sarcoma
F09 01	Organic mental disorder of brain	Z51.801	Immunotherapy for neoplasm
G96.103	spinal meningeal cyst, unspecified	Z51.802	Symptomatic

I69.801	Sequelae of cerebrovascular disease, unspecified	Z90.003	treatment for neoplasm Acquired absence of skull
K22.902	Neoplasm of oesophagus,unspecified	Z90.402	Acquired absence of stomach
K63.901	Neoplasm of colon,unspecified	Z93.301	Colostomy status
K86.901	Pancrea space-occupying lesion,unspecified	Z95.002	Status following coronary stent implantation
M51.302	Intervertebral disc degeneration	Z96.601	Status following artificial joint replacement
M84.491	Pathological fracture, not elsewhere classified	Z98.818	Postsurgical states of malignant neoplasm of brain
N63 01	Lump in breast, unspecified	Z98.820	Postsurgical states of malignant neoplasm of prostate

Figures

Codes	Diagnosis descriptions (English) Official descriptions (Chinese)	Real diagnosis descriptions	RegExps
A16.501	Tuberculous exadative pleurisy 结核性渗出性胸膜炎	右侧结核性渗出性胸膜炎，左侧结核性胸膜炎，结核性渗出性胸膜炎，双侧结核性渗出性胸膜炎，右侧渗出性结核性胸膜炎	^(.){0,3}(左 右 双)?(侧)?结核(性)?渗出(性 型)胸膜炎 ^.(.){0,3}(左 右 双)?(侧)?渗出(性 型)结核(性)?胸膜炎
B02.901	Zoster without complication 带状疱疹	右侧胸背部带状疱疹，带状疱疹，左侧面部带状疱疹，带状疱疹（左腰部），带状疱疹（右侧头面部），带状疱疹（右足踝），左侧头颈带状疱疹，带状疱疹（左耳部）等	^(.){0,1}(左 右)?(上 下)?(侧)?(头 颈 胸 背 腰 腹 肢 面)?(头 颈 胸 背 腰 腹 肢 面)?(部)?带状疱疹^(.){0,3}\$ ^.(.){0,1}(左 右)?(上 下)?(侧)?(头 颈 胸 背 腰 腹 肢 面)?(头 颈 胸 背 腰 腹 肢 面)?(部)?带状疱疹((\ \ < \ >)([^\])*(\) \ > \ >)(.){0,3}\$
D50.001	Hemorrhagic anemia 失血性贫血	重度失血性贫血，慢性失血性贫血，失血性贫血	^(.){0,3}(轻 中 重 急 慢)?(度 性)?失血性贫血
E02 01	Subclinical hypothyroidism 亚临床甲状腺机能减退	亚临床型甲减，亚临床性甲减，亚临床甲减，亚临床甲状腺机能减退症，亚临床甲状腺功能低下等	^(.){0,3}亚临床(型 性)?甲(状腺)?(机 功)?(能)?减(退 低)?(症)?
F06.702	Cognitive disorder 认知功能障碍	中度认知功能障碍，血管性认知功能障碍，老年性认知功能障碍，认知功能下降，认知功能减退等	^(.){0,1}(老年 血管 中 重)?(度 性 期)?认知(功能)?(障碍 下降 减退)^.(.){0,3}\$ ^.(.){0,3}认知(功能)?(障碍 下降 减退)((\ \ < \ >)([^\])*(\) \ > \ >)(.){0,3}(老年 血管 中 重)(性)?
G93.002	Arachnoid cyst 蛛网膜囊肿	蛛网膜囊肿，右侧颞部蛛网膜囊肿，颅内蛛网膜囊肿，左侧枕部蛛网膜囊肿等	^(.){0,3}(左 右 双 后)?(侧)?(颞 枕 顶 极 部 叶)?(颅 窝 内)?蛛网膜囊肿
H33.508	Choroidal detachment associated with retinal detachment 脉络膜脱离型视网膜脱离	右眼脉络膜型视网膜脱离，左眼脉络膜脱离型视网膜脱离，右眼脉络膜型视网膜脱离	^(.){0,3}(左 右 双)?(眼)?脉(络膜)?脱(离)?(型 性)(视)?网(膜)?脱(离)?
I25.101	Coronary atherosclerotic heart disease 冠状动脉粥样硬化性心脏病	冠心病，冠状动脉性心脏病，冠状动脉粥样硬化性心脏病，冠状动脉硬化性心脏病，冠状动脉心脏病	^(.){0,3}冠状动脉^(.){0,8}(性 型)心脏病 ^.(.){0,3}冠心病
O35.901	Fetal malformation 胎儿畸形	胎儿畸形，孕25周G2PO胎儿畸形，中孕胎儿畸形	^胎儿畸形 ^.(孕(.*周)(.){0,2})胎儿畸形 ^.(中 晚)(期)*孕^(.){0,2})胎儿畸形
R10.401	Abdominal pain 腹痛	腹痛待查，腹痛待诊，腹痛，右下腹痛，左侧腹痛待查	^(.){0,3}(左 右 双)?(上 中 下)?(侧)?腹痛(原因)?(待查 待诊)?

Figure 1

Examples of Diagnosis Descriptions Transformation into RegExps. The table shows the description models of the regexps were constructed by special symbols joining diagnosis descriptions. A diagnosis can be described by many different Chinese words, so we use Chinese to better present the application of regular symbols. This modeling method is also applicable to other languages.

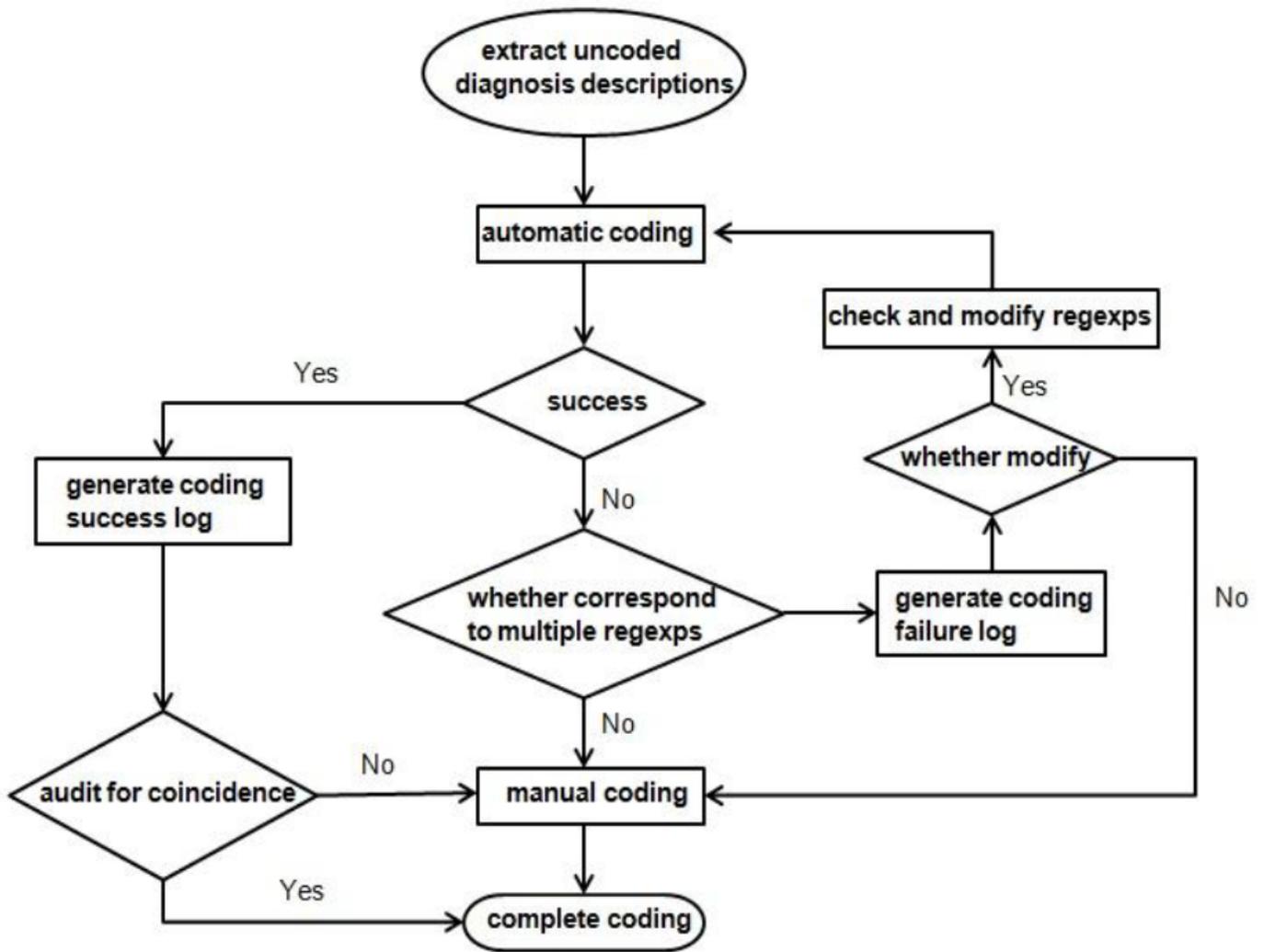


Figure 2

The flow chart of automatic coding with ICD-10. The process from waiting for coding to complete coding.

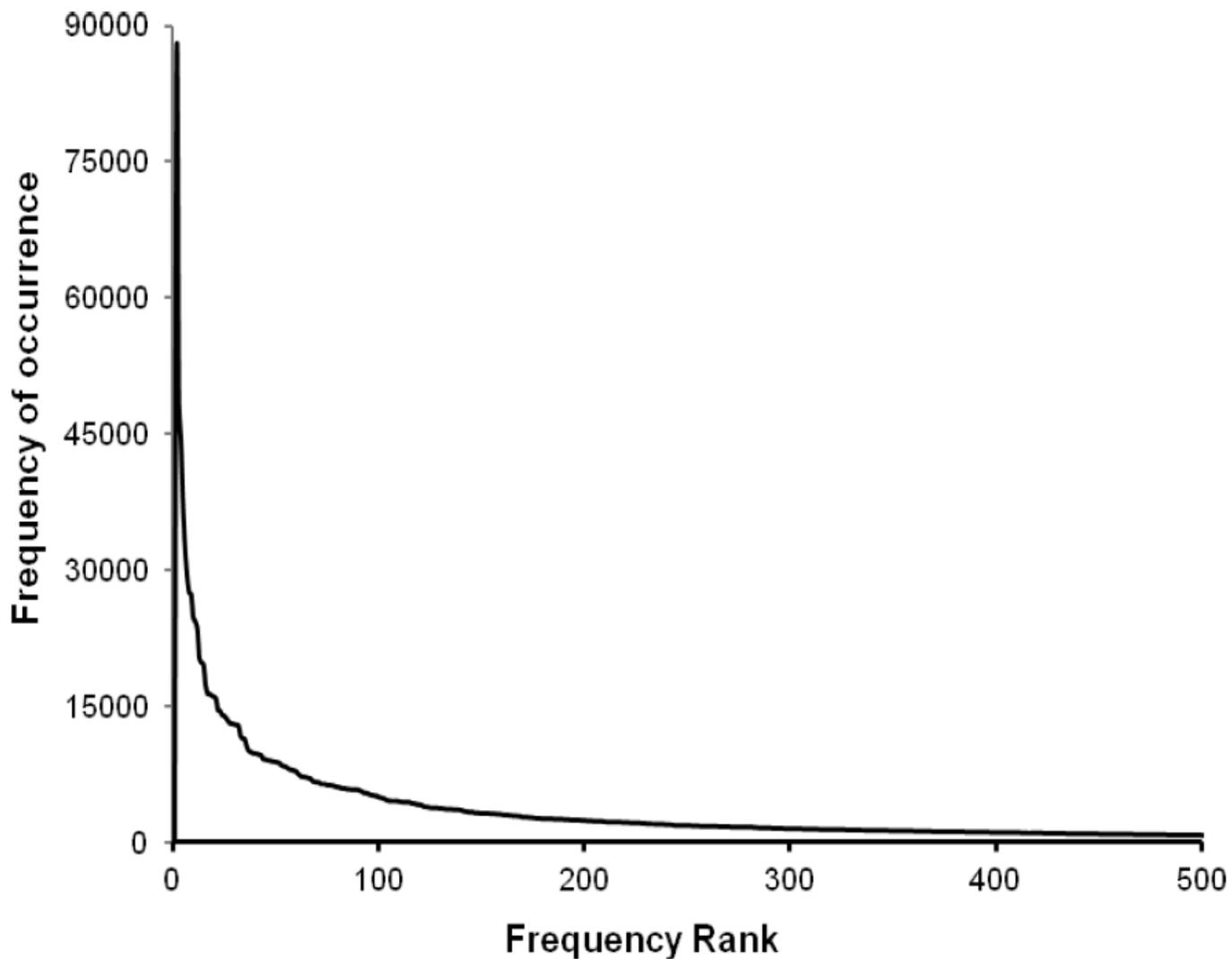


Figure 3

Distribution of the top 500 diagnosis codes. Horizontal and longitudinal ordinate represent the frequency rank and the frequency of occurrence, respectively, from January 1, 2014 to June 30, 2017 .

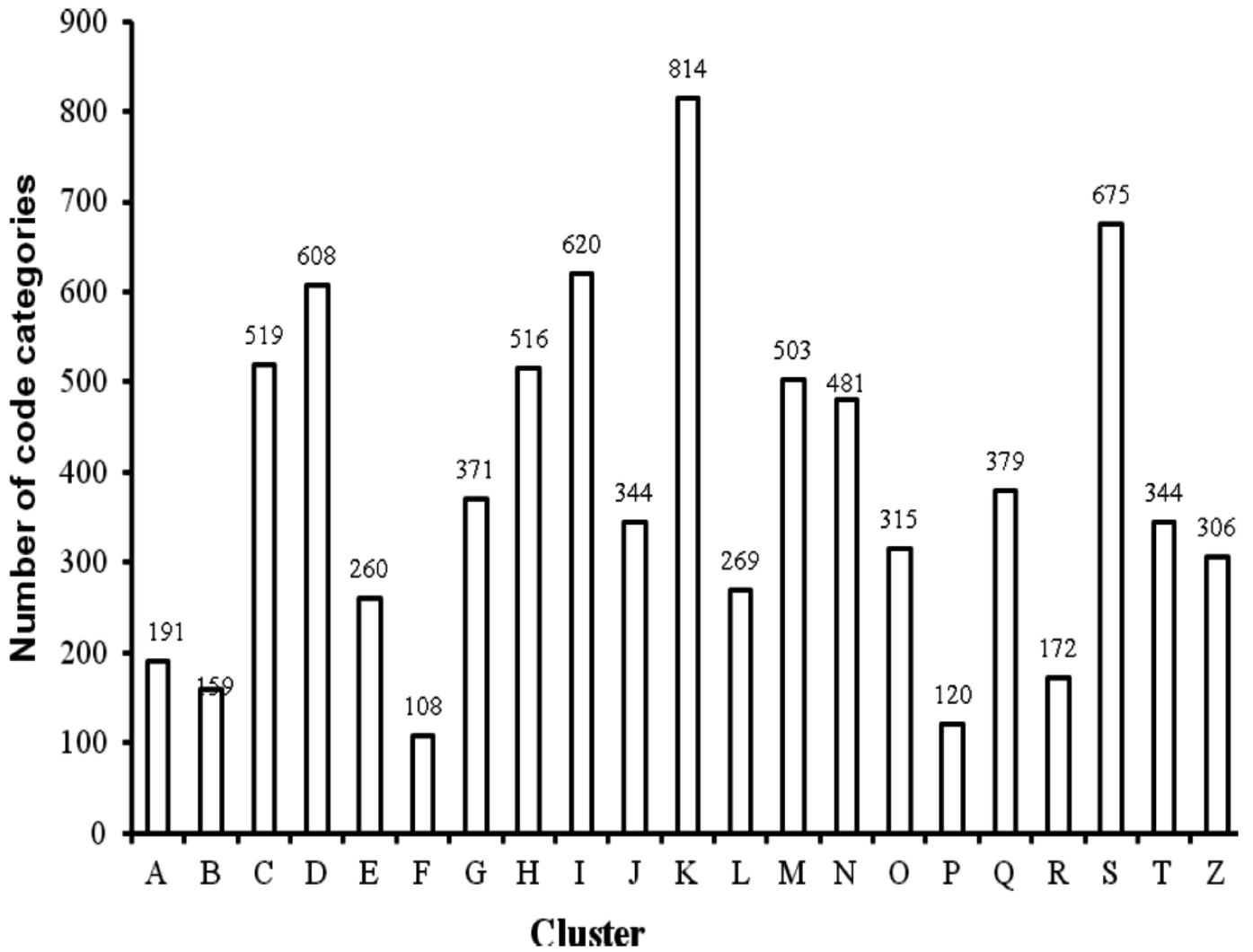


Figure 4

Number of code categories per cluster. Horizontal ordinate represents the each chapter of disease coding according to ICD-10. Longitudinal ordinate represents the number of code categories corresponding to each cluster from January 1, 2014 to June 30, 2017.

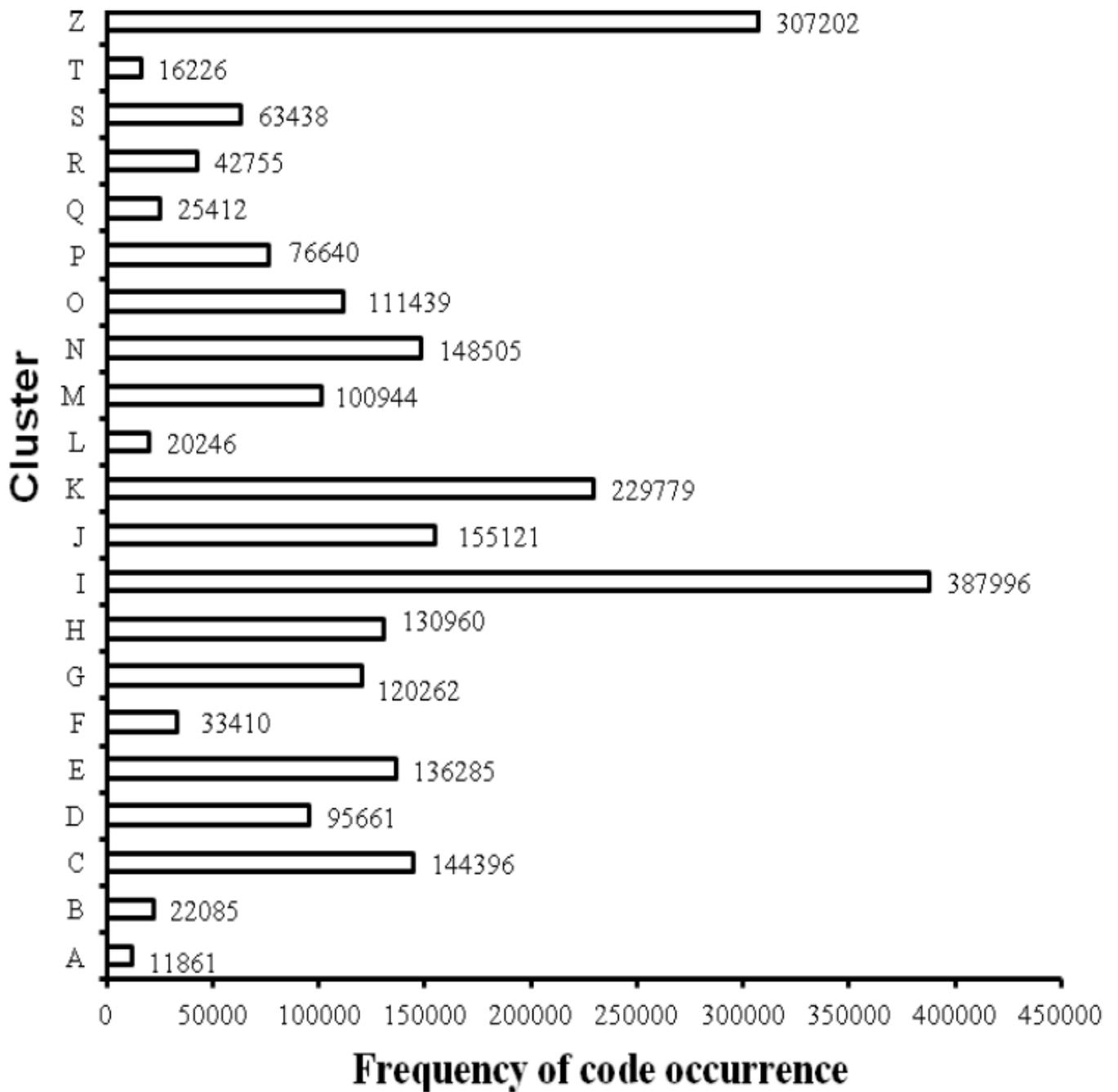


Figure 5

Number of diagnosis codes per cluster. Horizontal ordinate shows the frequency of code occurrence corresponding to each cluster from January 1, 2014 to June 30, 2017.

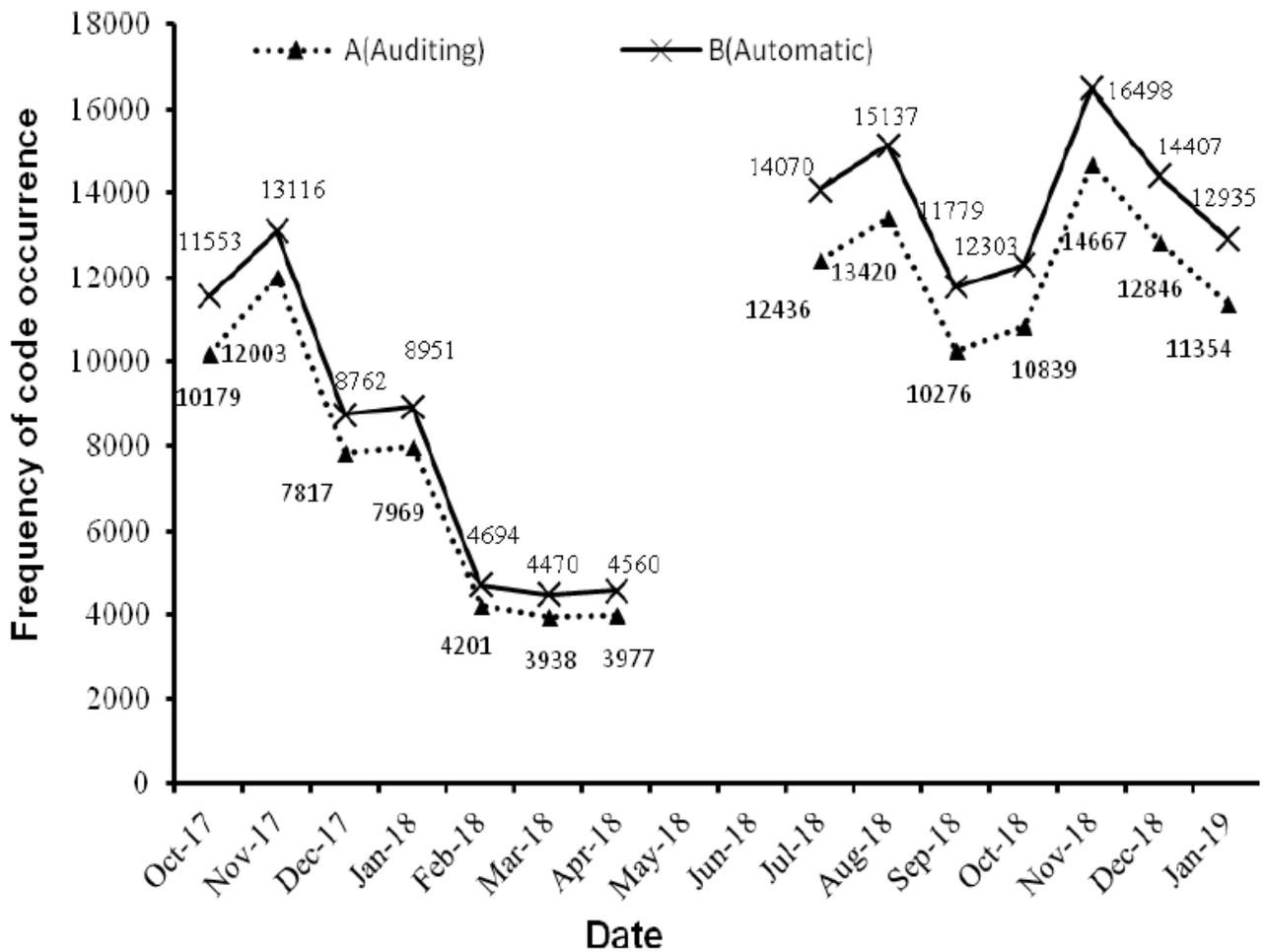


Figure 6

Comparison between the number of correct auditing codes and the automatic codes. Longitudinal ordinate shows the coding quantity per month from October 1, 2017 to April 30, 2018 and from July 1, 2018 to January 31, 2019. Curve A and B represent the correctly assigned diagnosis codes by the code auditors and the automatically completed codes by the coding system, respectively.

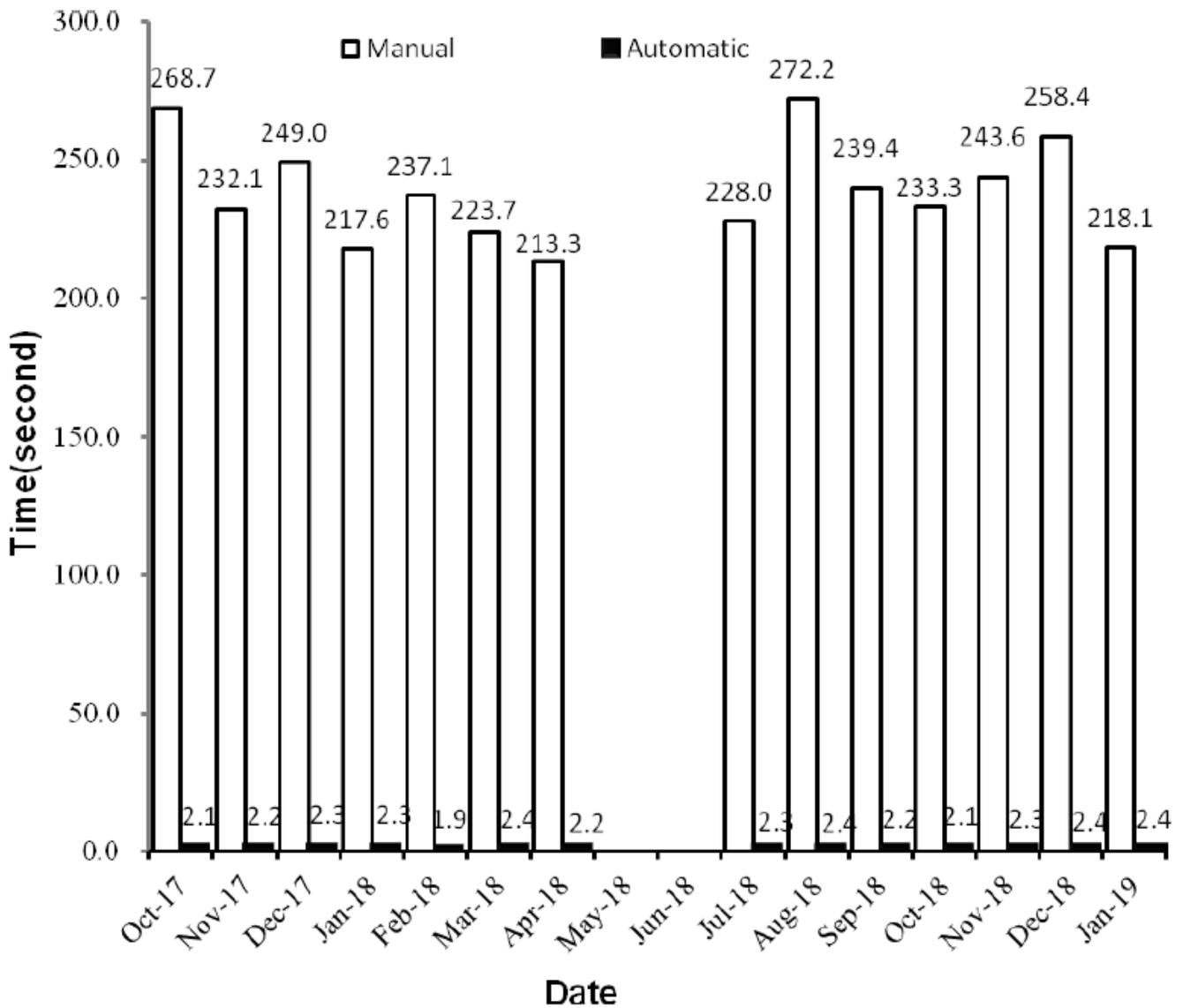


Figure 7

Time needed for manual coding and automatic coding. Longitudinal ordinate shows the average time-consuming (seconds) per ten codes every month from October 1, 2017 to April 30, 2018 and from July 1, 2018 to January 31, 2019.

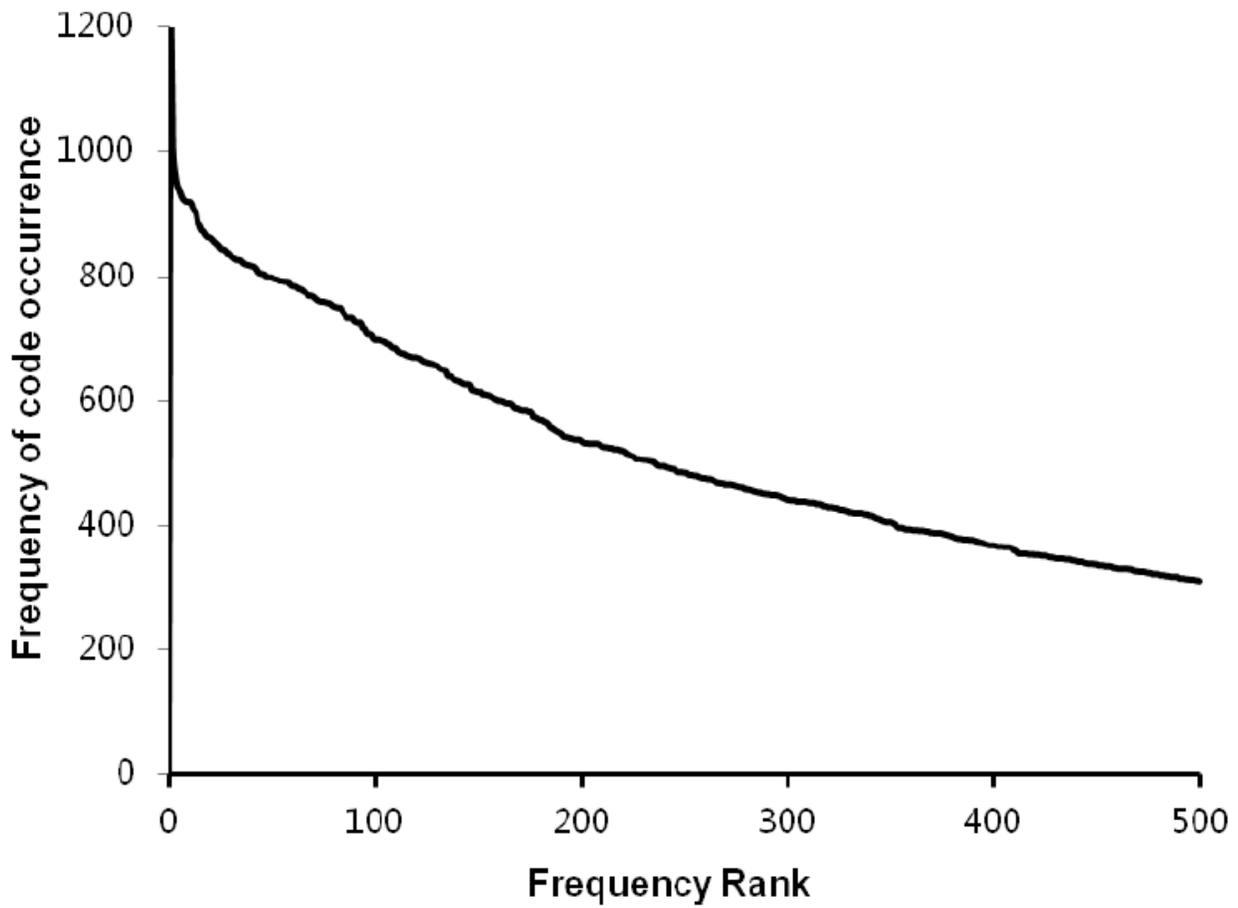


Figure 8

Distribution of the 500 high-frequency codes of the top 1000. Horizontal shows the other unmatched (in the first stage) 500 codes of the top 1000 from January 1, 2014 to April 30, 2018 .

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Proceduralcodes.doc](#)
- [Originaldata.xls](#)