MEDINFO 2017: Precision Healthcare through Informatics
A.V. Gundlapalli et al. (Eds.)

doi:10.3233/978-1-61499-830-3-427

A Hybrid Method for ICD-10 Auto-Coding of Chinese Diagnoses

Zheng Jia[a], Weifeng Qin[a], Huilong Duan[a], Xudong Lv[a], Haomin Li[b,c]

[a] College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou, China
[b] The Children's Hospital, Zhejiang University, Hangzhou, China
[c] The Institute of Translational Medicine, Zhejiang University, Hangzhou, China

Abstract

The Chinese Version of Classification and Codes of Diseases (CCD) is an expanded version of ICD-10. Hospitals are required to assign CCD codes to discharge diagnoses in China. To handle the contradiction between a shortage of skilled CCD coders and increasing coding efficiency, a CCD auto-coding method is urgently needed. In this study a hybrid auto-coding method was proposed based on the lexical characteristics obtained through the analysis of a corpus of 1537 diagnoses with normative CCD code. It combines the rule-based approach, the Chinese characters-based distributed semantic similarity and the dictionary-based approach. The rule-based approach was proved to be efficient and precise at the cost of time and manpower. The semantic similarity approach shows poor performance. The old-fashioned dictionary-based approach ends in leading significance. The final accuracy of this hybrid approach is 96.9% in the test.

Keywords:

International Classification of Diseases; Clinical Coding; China

Introduction

The International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10) [1] was defined as a system of categories to which morbid entities are assigned according to established criteria. The ICD was used to translate diagnoses of diseases and other health issues from linguistic words into alphanumeric codes, which permits easy storage and retrieval, and systematic recording analysis of mortality and morbidity data. The classification consists of 22 chapters, which were divided into homogeneous blocks of categories, which are further subdivided to at most 10 subcategories each.

The Chinese Version of Classification and Codes of Diseases (GB/T 14396-2001) [2], short for CCD, is an expanded version of ICD-10 and is widely used in China. It has a strict and organized classification hierarchy expanded from four-char ICD-10 (9505 subcategory) to six-char codes (23106 expansion) that allows one-to-one mapping from diseases to codes to support the management, inquiry, search and statistical analysis of data.

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 discharge diagnoses in the summary page of medical records after January 1, 2012 [3].

Accurate CCD diagnosis coding is critical to patient care, billing purposes, and research endeavors [4]. Correct, standard and complete hospitalization diagnosis by doctors and careful implementation of coding principle and careful reading about medical record content are the guarantee of ensuring correct coding [5]. The patient's course of disease is defined as a procedure from admission to discharge with a narrative discharge summary including the primary and secondary diagnoses by manual input. The paper trail is defined as a creation of medical record from the recording of the admitting diagnosis to the assignment of the ICD codes after discharge [6][6].

Errors such as "upcoding" and misspecification may be introduced during the conding by the lacking of medical knowledge and coder's experience, attention, and persistence, the ability to interpret the diagnoses and the inconsistency between CCD phraseology and clinical phraseology, such as the synonyms and abbreviations used to describe the same condition. [6].

In most of large hospitals, several professional coders are hired by the medical record department to supplement the CCD codes in the summary page of hospital records [7]. Nevertheless, the contradiction between a shortage of skilled coders and an excess of hospital patients makes it a tough task to encode both the primary and the secondary discharge diagnoses with ICD-10. In a first-class hospital in China only two coders are hired to assign CCD codes for about 2000 inpatients everyday.

Physicians in the same department usually follow the same naming convention so that there is little difference in the record names of the same disease in its summary page of hospital records. Similar grammar mistakes and medical language process makes it feasible to analyze the underlying rules of encoding diagnoses into CCD codes and develop a CCD auto-coding tool for physicians.

To reduce the workload of human coders, Medori and Fairon [8] extracted necessary information from French discharge summaries and combined the symbolic approach and statistical approach, since a large corpus of clinical notes was available. Subotin and Davis [9] developed a system for predicting ICD-10-PCS codes from the English clinical narrative using partial hierarchical classification and confidence calculation and estimation. In China, Ning and Yu [10] proposed an algorithm to implement ICD-10 coding automatically for clinical diagnoses in Chinese and calculated the semantic similarity between terms by the definition of distributed semantic similarity. Considering the linguistic features of Chinese, the results indicate that term vectors constructed from words have a higher precision than that from Chinese characters.

In this paper, we analyze and summarize the characteristics of discharge diagnoses written by doctors in the Nephrology Department, Dayi Hospital. A rule-based data pre-processing method and a computer-assisted CCD coding method are proposed.

Methods

In this study, we collected a corpus, analyzed the lexical characteristics of Chinese diagnoses and proposed three approaches to auto-code CCD.

Corpus Collection

The corpus of patient Electronic Medical Record (EMR) summary pages were collected from a Clinical Data Repository (CDR) [11] system, which was implemented in the 2000-beded hospital in China. The summary page was in XML format and each summary page recorded one primary discharge diagnosis of the patient and its associated CCD code assigned by human coders. A corpus of 1537 diagnoses with CCD codes considered as the gold standard was randomly selected and parsed from data repository as the typical language samples to analyze the lexical characteristics.

Lexical Characteristics Analysis

Due to individual language habit, doctors may use diverse names to define the primary diagnoses and describe the status of disease [6]. ICTCLAS [12], which is a Chinese lexical analyzer, is used to tokenize the long disease names into minimal semantic units (Table 1).

Table 1 – Result of Tokenization

Original Name	Tokenization
慢性肾衰竭急性加重 chronic kidney failure exacerbate acutely	慢性肾衰竭/n 急性/b 加重/n
轻度系膜增生性IgA肾病 mild membranoproliferative IgA nephropathy	轻度/d 系膜增生性IgA肾病/n
无症状性血尿 asymptomatic haematuria	无/v 症状/n 性/ng 血尿/n
不典型膜性肾病 atypical membranous nephropathy	不/d 典型/a 膜性肾病/n

Tags: n – Noun, b - distinguishing words, v – verb, d – adverb, a – adjective, ng - Noun morpheme.

The analysis results include the tokenization result and observational conclusions (Table 2). The hand-input phrases usually consist of three components: multiple attributes, one core noun and punctations. The punctations consist of redundancy such as bracket and digit symbols. Several common wrongly written characters of discharge diagnoses are discriminated and the reason is analysed.

Table 2 – Semantic Class of Chinese Diagnoses

Semantic class	Examples
Attribute	*弥漫性* *膜性* 肾小球肾炎 *diffuse* *membranous* glomerulonephritis
Noun	弥漫性膜性*肾小球肾炎* diffuse membranous *glomerulonephritis*
Punctation	肾炎综合征，膜性肾病 nephrotic syndrome, membranous nephropathy 肾炎综合征（膜性肾病） nephrotic syndrome (membranous nephropathy)
Errors	homophone, Chinese character in similar form, synonymous symbol, synonymous words, et al.

Computer-assisted Coding Method

We proposed three approaches for automated CCD code assignment: the rule-based approach, the Chinese characters-based distributed semantic similarity approach and dictionary-based approach.

Rule-based auto-coding

The rule-based approach contains nine irregular naming patterns established from practice and experience (Table 3).

Superfluous prefix or suffix are omitted. Typical prefix words include anatomic parts(eg. "下肢"Chinese for "leg"), nouns of locality(eg. "左" Chinese for "left") and scope(eg. "多发" Chinese for "multiple"), words end with particular Chinese characters(eg. 性 property).

Table 3 – Rules for CCD Auto-coding

Pattern	Original Name	Processed Name
prefix term	**右**肾结核 **right** kidney tuberculosis	肾结核 kidney tuberculosis
term **suffix**	腰椎管狭窄**症** lumbar spinal stenosis **disease**	腰椎管狭窄 lumbar spinal stenosis
term1 **and/with** term2	肾病综合征**合并**急性肾损伤 nephrotic syndrome **and** acute injury of kidney	肾病综合征 nephrotic syndrome 急性肾损伤 acute injury of kidney
term1 **punctation1** term2 **punctation2**	肾炎综合征**，**膜性肾病 nephrotic syndrome, membranous nephropathy	肾炎综合征 nephrotic syndrome 膜性肾病 membranous nephropathy
character in similar form	肾病综合**症** nephrotic syndrome	肾病综合**征** nephrotic syndrome
homophone	急进**型**肾炎综合征 rapidly progressive nephritic syndrome	急进**性**肾炎综合征 rapidly progressive nephritic syndrome
synonymous symbol	Roman numeral ⇔Arabic numeral	
synonymous words	甲状腺**功能**减退症 hypothyroidism	甲状腺**机能**减退症 hypothyroidism
remove non-Chinese character	腰5骶1椎间盘突出 lumbar 5 sacrum 1 disc herniation	腰骶椎间盘突出 lumbar sacrum disc herniation

Conjunctions (eg. “合并” Chinese for “accompanied”) and meaningful punctations(eg. “逗号” Chinese for “comma”) are sometimes used to connect two individual disease names. The CCD dictionary may exclude the combination but include the individual.

Each permutation of the rules applied to one original diagnosis name will generate a candidate. The candidates that can be matched to any disease name in CCD dictionary or custom dictionary (much more on this later) completely will be selected. The standard disease name will also be processed if it contains particular characters or symbols. The rule-based auto-coding is invalid if it’s not an exact match. The candidate with a perfect match and maximum character count will be recommended as the paired CCD code.

Chinese characters-based distributed semantic similarity

The Chinese characters-based distributed semantic similarity [10] is a method which can implement ICD-10 coding automatically for clinical diagnoses in Chinese and has a high precision in the test set. It is measured by the cosine similarity between vectors converted from Chinese characters.

$\vec{p} = (p_1, p_2, \cdots p_M)$ and $\vec{q} = (q_1, q_2, \cdots q_N)$ are defined as the standard CCD name (the count of Chiniese characters is m) and the hand-input name (the count is n) repectively.

$\overrightarrow{p'} = (p_{i_1}, p_{i_2}, \cdots p_{i_K})$, $\overrightarrow{q'} = (q_{j_1}, q_{j_2}, \cdots q_{j_K})$ are the vectors extracted from $\vec{p}$ and $\vec{q}$ where $p_{i_k} = q_{j_k} (k \in [1, K])$, $\overrightarrow{p'} = \overrightarrow{q'}$, $p_i \in \{p_1, p_2, \cdots p_m\}$, $q_j \in \{q_1, q_2, \cdots q_n\}$. The similarity score is

$$\mathrm{sim}(\mathrm{p}, q) = \frac{K}{M} \cdot \frac{K}{N} \cdot |\vec{\imath} - \vec{\jmath}|$$

where

$$\vec{I} = (i_1, i_2, \cdots i_K), i_k \in [1, M]$$
$$\vec{J} = (j_1, j_2, \cdots j_K), j_k \in [1, N]$$

and M is the count of Chiniese characters of the standard CCD name, N is the count of Chiniese characters of the hand-input name, K is the count of common characters.

For example, there are one original disease name 巨幼细胞贫血(megaloblastic anaemia) and three candidates of CCD name 巨幼红细胞性贫血(megaloblastic anaemia)，巨幼细胞遗传性贫血(megaloblastic hereditary anaemia), 营养性巨幼细胞性贫血(nutritional megaloblastic anaemia). We have

$$\vec{q} = (巨, 幼, 细, 胞, 贫, 血), N = 6$$
$$\overrightarrow{p_1} = (巨, 幼, 红, 细, 胞, 性, 贫, 血), M_1 = 8$$
$$\overrightarrow{p_2} = (巨, 幼, 细, 胞, 遗, 传, 性, 贫, 血), M_2 = 9$$
$$\overrightarrow{p_3} = (营, 养, 性, 巨, 幼, 细, 胞, 性, 贫, 血), M_3 = 10$$

The extracted vectors are $\overrightarrow{p_1'} = \overrightarrow{p_2'} = \overrightarrow{p_3'} = \overrightarrow{q'} = (巨, 幼, 细, 胞, 贫, 血)$, $K = 6$, where the sequence vectors and similarity scores are

$$\vec{J} = (1,2,3,4,5,6)$$
$$\overrightarrow{I_1} = (1,2,4,5,7,8), \quad \mathrm{sim}(p_1, q) \approx 2.37$$
$$\overrightarrow{I_2} = (1,2,3,4,8,9), \quad \mathrm{sim}(p_2, q) \approx 2.83$$
$$\overrightarrow{I_3} = (4,5,6,7,9,10), \quad \mathrm{sim}(p_1, q) \approx 4.95$$

The non-Chinese characters are removed before similarity calculation. The candidate code with the lowest similarity score will be assigned to the disease name. In the above example, the CCD code of p_1 should be assigned to q.

Dictionary-based approach

The dictionary-based approach is to construct a custom dictionary which extended the CCD standard disease name with disease names used by doctors and its corresponding CCD codes given by human coders. The scope of the dictionary tends to expand over time. The term absented from CCD dictionary will be assigned to its newly assigned CCD code when it appears again.

The three approaches are integrated to follow the priority order: custom dictionary > rule-based auto-coding > Chinese characters-based distributed semantic similarity. If the original diagnosis fails to match any CCD of the standard CCD dictionary or the custom CCD, it is then processed by the rules. This procedure may generate several candidates. If no candidate is obtained, the auto-coding fails. Otherwise the candidates are compared with CCD of the standard CCD dictionary or the custom CCD. If no candidate matches successfully, the CCD code with a highest score of Chinese characters-based distributed semantic similarity is selected as the final result. In a situation where there are multiple matches in standard or extended CCD dictionary, the first candidate in the default sort is selected as the final match. The flow diagram of the hybrid method is shown in Figure 1.

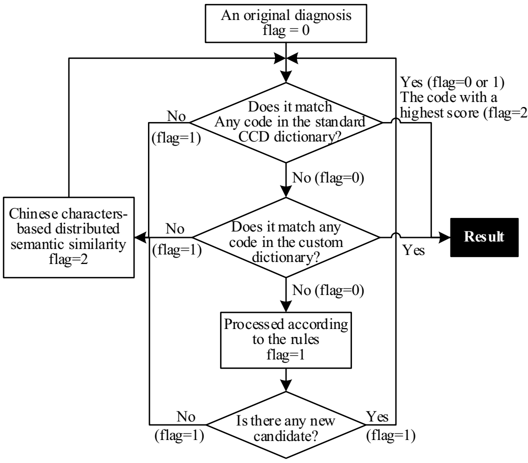

Figure 1 – The Flow Diagram of the Hybrid Method

Results

The corpus of 1537 diagnoses were auto-coded according to the hybrid approach. Then we verified the validity of recommended CCD codes by the gold standard codes assigned by human coders. In Table 4 the count of right results of each pattern was calculated programmatically. Each pattern contributes to a bit of the final result and the percentage number indicated the contribution.

About ten percent of raw diagnoses written by doctors could be mapped to one CCD code without any medical language processing. It indicated that the language habit and common parlance in the clinical context varied among terms in CCD dictionary.

Rule-based auto-coding was an effective method with high accuracy but low recall. We applied regex functions for specified patterns; that is, candidates were searched among both the standard disease names and the original disease names according to the rules. It supplemented 42.4% diagnoses with correct CCD codes. We developed and reviewed every rule carefully at the cost of time and manpower. A further investigation was carried out and it came to the conclusion that the auto-coding rules were largely consistent across all departments.

Table 4 – Auto-coding Result

Approach	Pattern	Count	Percent
Direct Match		161	10.5%
Processed Standard Name	term1 punctuation1 term2 punctuation2	285	42.4%
	prefix term suffix	12	
Processed Original Name	prefix term	114	
	term suffix	27	
	term1 and/with term2	4	
	term1 punctuation1 term2 punctuation2	123	
	character in similar, form, homophone, synonymous words	18	
	synonymous symbol	57	
	remove non-Chinese character	12	
Chinese characters-based distributed semantic similarity		18	1.2%
Dictionary-based		659	42.9%
Mismatch		47	3.1%
Count		1537	100%

Chinese characters-based distributed semantic similarity method was a method dedicated to Chinese characters with mediocre accuracy, poor efficiency and small influence. Only 1.2% discharge diagnoses were under valuable influence.

The dictionary-based approach improved the matching rate from 52.9% to 97%. It proved to be an invalid approach initially and would play a leading role as it scaled. It may start out as unreliable, but it can be trained and developed into a practical handbook coders can rely on. At present, most of the hospitals seldom established this kind of dictionary and even had no information system assisting the human coders [13].

For mismatched cases, some errors are analyzed. One case is that the original name is 双下肢动脉硬化伴多发斑块形成 (the hardening of the arteries of both lower limbs with multiple plaque buildup) and the standard name is 下肢动脉动脉粥样硬化(atherosclerosis of arteries of lower limbs). Another case is that the original name is 系统性小血管炎肾损害(kidney damage due to (not mentioned in Chinese) systemic small-vessel vasculitis) and the standard names are 系统性血管炎 (systemic small-vessel vasculitis) and 系统性结缔组织疾患引起的肾小球疾患(glomerular disorders in systemic connective tissue disorders, ICD-10: N08.5*). It infers that mismatches may be caused by the usage of totally different syntax.

Discussion

The combination of the three approaches reached 96.9% precision for the auto-coding of 6-char CCD in the Nephrology Department. Ning et al.[14] used a hierarchical method to automatically encode Chinese diagnoses through semantic similarity estimation, reaching the precision of 4-char coding is about 92%. It indicates that our method improved the fineness without diminution of accuracy.

The rule-based method and the dictionary-based approach cost running times within 1 second, while the Chinese characters-based distributed semantic similarity is long-running process. The average running time of the method proposed by Ning[14] is about 1 second. For the little sense made by the Chinese characters-based distributed semantic similarity method, it is suggested to be abandoned.

The rule-based method in this paper is essentially an extended dictionary matching approach enriched with rule pre-processing. It is not a traditional rule based approach for other tasks. A hospital-wide, highly efficient, precise auto-coding system coordinating the rule-based method and the dictionary-based approach is the next challenging task in our project.

The rule-based method and dictionary-based approach take merely a fraction of operation time. However, the Chinese characters-based distributed semantic similarity method requires more internal storage and longer execution time, because the dictionary is large and the similarity calculation and comparison between each code is time-consuming.

Conclusion

In this paper we introduced the usage condition of ICD-10 in China and the expanded version of ICD-10 known as CCD. The existing approaches assisting human coders to assign CCD codes to diagnoses context were reviewed. Three methods were proposed. Rule-based auto-coding method is of high effectiveness and precision. Chinese characters-based distributed semantic similarity method performed poorly. The dictionary-based approach proved to be gradually crucial and reliable.

Howerver, there are still limitations. The hybrid appoach is of low portability and interoperability, because rules for auto-coding and the custom dictionary differ from hospital to hospital. Rule-making and dictionary construction for each single hospital is of high cost and low efficiency. In future work, we will conduct the in-depth analysis of auto-coding and develop a hospital-wide system to assist the work of human coders.

Acknowledgements

This research was financially supported by the national key research project 2016YFC0901905.

References

[1] W.H. Organization, *International statistical classification of diseases and health related problems (The) ICD-10*, World Health Organization, 2004.

[2] General Administration of Quality Supervision, Inspection and Quarantine of the People's Republic of China, Classificatoin and Codes of Diseases, in, Standards Press of China, 2002.

[3] Notice of the General Office of the Ministry of Health on the Issues concerning Classification and Codes of Diseases (Revision), 2011.

[4] C.P. Chen, S. Braunstein, M. Mourad, I.-C.J. Hsu, D. Haas-Kogan, M. Roach, and S.E. Fogh, Quality improvement of International Classification of Diseases, 9th revision, diagnosis coding in radiation oncology: Single-institution prospective study at University of California, San Francisco, *Practical radiation oncology* **5** (2015), e45-e51.

[5] C. Li, X. Chen, An Analysi s on Reasons of Wrong Coding for 5000 Cases of Medical Records, *Chinese Medical Records* **12** (2011), 23-24.

[6] K.J. Omalley, K.F. Cook, M.D. Price, K.R. Wildes, J.F. Hurdle, and C.M. Ashton, Measuring Diagnoses: ICD Code Accuracy, *Health Services Research* **40** (2005), 1620-1639.

[7] Y. Wang, N. Guo, X. Bai, Discussion on Coder Cultivation in County-Level Hospitals, *Chinese Medical Record*, ***12***(2015):81-82

[8] J. Medori and C. Fairon, Machine learning and features selection for semi-automatic ICD-9-CM encoding, in: *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*, Association for Computational Linguistics, 2010, 84-89.

[9] M. Subotin and R. Davis, A system for predicting ICD-10-PCS codes from electronic health records, *ACL 2014* (2014), 59.

[10] X. Ning, M. Yu, Algorithmic Research on Automatic Coding of Clinical Diagnoses Based on Semantic Similarity Calculation, *Journal of Medical Informatics*, **37** (2016), 52-56.

[11] L. Min, L. Wang, X. Lu, and H. Duan, Case Study: Applying OpenEHR Archetypes to a Clinical Data Repository in a Chinese Hospital, in: *MEDINFO 2015: EHealth-enabled Health: Proceedings of the 15th World Congress on Health and Biomedical Informatics*, IOS Press, 2015, p. 207.

[12] P. Zhang, K. Yu, Y. Xiong, and Q. Liu, HHMM-based Chinese lexical analyzer ICTCLAS, in: *Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17*, Association for Computational Linguistics, 2003, 184-187.

[13] L. Yang, M. Yu, An Application of a Computer Aided ICD-10 Coding System, *Chinese Medical Record*, (2015), 28-32.

[14] W. Ning, M. Yu, and R. Zhang, A hierarchical method to automatically encode Chinese diagnoses through semantic similarity estimation, *BMC Medical Informatics and Decision Making* **16.1** (2016), 30

Address for correspondence

Zheng Jia
College of Biomedical Engineering and Instrument Science,
Zhejiang University, Hangzhou, China.
Email: jiaz@vico-lab.com
Tel: +8615858296829