MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220041

A Semi-Automatic Data Cleaning & Coding Tool for Chinese Clinical Data Standardization

Yani Chen[a], Qi Tian[a], Hailing Cai[a], Xudong Lu[a,b*]

[a] *College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou, Zhejiang Province, China*
[b] *Country Two Key Laboratory for Biomedical Engineering, Ministry of Education, y, Hangzhou, Zhejiang Province, China*

Abstract

The clinical data often have limited usefulness because of the diversified expression. Chinese clinical data standardization can improve the usability of clinical data. The complexity of data cleaning and coding for Chinese clinical data prompted the turn of low-effective manual coding into the computer-aided tool. This study established the universal data cleaning and coding process and tool for Chinese clinical data standardization, which can greatly improve human efficiency. The process included the preprocessing, text similarity algorithm, and manual review. The standardization process proved effective for the diagnosis, drug, and examination data standardization task and can be used gradually in other clinical domains. The semi-automatic data cleaning and coding can reduce the half time for standardization, and it was used in hospitals in Beijing.

Keywords:
Standardization; Automatic Data Processing

Introduction

Millions of clinical data are now routinely collected across diverse hospitals in China. The utility of clinical data often has limited because of the diversified expression of clinical texts wrote by different doctors. Many studies contribute to moving the format of clinical data to a structured format[1-3], but the fundamental barrier to analyzing clinical data is a gap in the standardization[4]. Standardizing clinical data with controlled or standardized vocabularies (terminologies) can improve the quality of clinical data and promote clinical data interoperable across different health care systems[5].

Standardization of Chinese clinical data is a complex and time-consuming process, usually performed by trained coders. First, the coders remove all duplicate data to reduce the workload. Second, the coders select keywords from the clinical data based on experience, then search and conform matching results from the standard dictionary. Because Chinese writing has more diverse expressions with a non-word boundary style that means no delimiter to separate a word in texts, unlike the English language. It may be necessary to perform multiple keyword changes and searches. In addition, coders should pay attention to filtering error data and manually delete it. Finally, the codes are matched with the original data. The amount of clinical data is growing rapidly in China, while the number of coders remains the same. It prompted the search for tools that assist manual standardization [6]. A semi-automatic coding tool can greatly improve efficiency and relieve the insufficient coders where the role of the coders would be to check and complete the codes provided by the tool[7].

The novelty and complexity of data cleaning & coding for Chinese clinical data present unprecedented challenges for the development of semi-automatic schemes and tools. There are many data quality problems with raw clinical data, and the processing required for clinical data in different domains has significant differences[8]. It is necessary to clean data before data coding. Some studies developed recommendation systems offering top k standardization results for given clinical data. An online standardization system, Clinical-Coder, aims to assign ICD codes to Chinese clinical notes with dilated convolutional attention network with an N-gram Matching mechanism[9]. Fei Teng et.al developed a cross-textual attentional ICD coding method and applied it in a computer-aid coding system[10]. Zhou L et.al use regular expressions (regexps) to establish a practical, automatic ICD-10 coding system between diagnosis descriptions[11]. The above tools can meet the standardization of clinical data to a certain extent but only target a single clinical domain. There is still a lack of a universal standardization process and tool. The general steps of data coding in different clinical domains can be summarized into a universal standardization process. The process can help develop a configurable visualization semi-automatic data cleaning & coding tool to accelerate the standardization in multiple clinical domains.

In view of the above situation, a universal process and a semi-automatic data cleaning & coding tool for Chinese clinical data standardization were developed, which can greatly improve human efficiency, and meanwhile accelerate the secondary use for clinical informatics.

Methods

The architecture of the standardization process is shown in Figure 1. The Chinese clinical data are cleaned by preprocessing the text. Then, an appropriate text similarity algorithm is selected to recommend the standardized results. The standardized Chinese clinical data are output after the manual review of the recommendation.

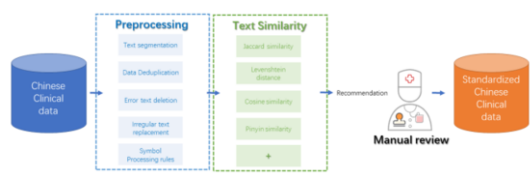

Figure 1– The architecture of the standardization process

Preprocessing

Data preprocessing includes five modules, text segmentation, data deduplication, error text deletion, irregular text replacement, and symbol processing rules. The semi-automatic tool used the five modules and supports manual expansion to preprocess raw clinical data shown in Figure 2.

Figure 2– The configurable preprocessing in the tool

Text segmentation

Chinese texts have no explicit word boundary markers. Text segmentation is intercepted by reading in human coding, the tool automatically splits the text through rules and adjusts error results manually with higher efficiency. Many highly robust rules to sectionize the clinical data into logical parts were developed. The segmentation delimiters include “;”, “,”, “@”, “.”, “\”, etc. (16 kinds in total). So we can divide a clinical text, “头晕;颈肌紧张、 脑膜瘤” into three parts, “头晕”, “颈肌紧张” and “脑膜瘤”, then standardize them separately.

Data Deduplication

Deduplication that eliminates duplicate entries of the same information is very sophisticated[12]. In the preprocessing process, only exact match deduplication was processed by the simple join operation in SQL or its equivalent. Compared with manual deduplication, the tool is more convenient and faster with a data deduplication module.

Error text deletion

The emergency and busyness for treating patients result in some low-quality error texts in clinical data[10]. Error texts usually have nothing to do with clinical information but are reminders or marks for doctors in the process of treating patients. Error texts have some meaningless special symbols and description, such as “°”, “▲”, “[甲]”([A]), “进口”(Import), “健康查体”(Health check), “病理待报”(Waiting for Pathology) and “复发可能”(Possible recurrence). Error text deletion needs reading and thinking carefully by experienced coders that is a time-consuming and inefficient stage in human standardization. A symbols and phrase dictionary was developed to filter and delete the error text and can be manually expanded in the process of standardization. Error text deletion will gradually become efficient and accurate in the tool with the improvement of the dictionary.

Irregular text replacement

The emergency and busyness for treating patients result in some abbreviations in clinical data[13; 14]. Irregular text replacement needs re-enter manually by experienced coders in human coding. Some description forms recognized to be incorrect were changed into normalization automatically through rules set that can be continuously supplemented and improved during the use of the semi-automatic tool. For example, "中耳炎右"(Otitis media R) can be automatically converted into a standardized "右侧中耳炎"(Right otitis media) through the rules set.

Symbol processing rules

The raw clinical data use symbols to supplement some details[15]. The details and symbols are needed to be revised by experienced coders in manual coding. The symbol processing rules were developed to fast handle different symbols, some examples of symbol processing were shown in Table 1.

Table 1– Examples of symbol processing

Symbol	Raw text	After processing
()	骨关节炎（多部位） Osteoarthritis (multiple parts)	多部位骨关节炎 Multi-site osteoarthritis
“”	左肩“脱位” Left shoulder "dislocation"	左肩脱位 Left shoulder dislocation
!	脑梗塞！ Cerebral infarction!	脑梗塞 Cerebral infarction
【】	1子宫肌瘤【多发性】 1Uterine fibroids [multiple]	多发性子宫肌瘤 Multiple uterine fibroids
{}	面部色素痣{激光治疗后} Pigmented facial nevus {after laser treatment}	面部色素痣;激光治疗后 Pigmented facial nevus;Laser treatment

Text similarity

Chinese clinical data were turned into Chinese clinical phrases through preprocessing. Then the text similarity algorithms were used to standardize Chinese clinical phrases shown in Figure 3. The text similarity algorithms implemented in the semi-automatic data cleaning & coding tool were introduced as following.

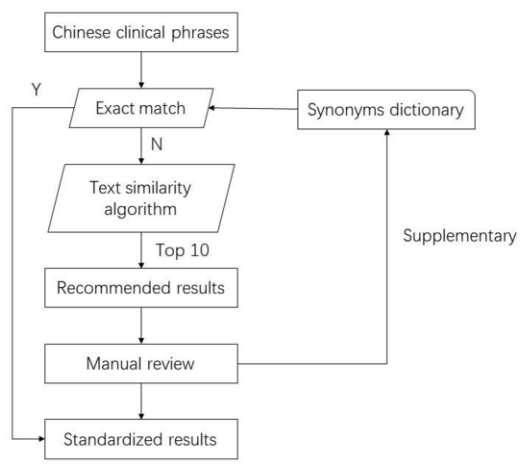

Figure 3– The standardization process of Chinese clinical phrases with the text similarity algorithm

Edit distance

The edit distance refers to the minimum number of insertions, deletions, and substitutions required to transform one string into the other[16].

Jaccard coefficient

The Jaccard coefficient refers to the intersection divided by the union of two texts[17].

Cosine text similarity

Cosine text similarity measures the cosine of the angle between two word vectors[18]. Representations of text such as bag-of-words models were used to turn the Chinese clinical phrases into word vectors.

Chinese Phonetic Similarity

Chinese phonetic similarity algorithms identify words and phrases with similar pronunciation[19]. For example, the pronunciation of "头痛" is "toutong" and the pronunciation of "头疼" is "touteng", which has a strong Chinese Phonetic Similarity.

Tongyici Cilin Similarity

Tongyici Cilin is a Chinese concept dictionary that is similar but small to WordNet[20]. Tongyici Cilin Similarity used the coding and structural characteristics of Tongyici Cilin to calculate the distance between two texts.

Manual review

It is difficult for the automatic coding algorithm to achieve 100% accuracy, so in practice, to a certain extent, the tool should also have a manual review[21]. The data covered by the synonyms dictionary or completely matched does not perform the manual review, which is an important step to improve efficiency. The semi-automatic tool can accelerate the standardization process by automatically recommending standardized results, and ensure the accuracy of standardization through manual review shown in Figure 4. The results of the manual review will be added to improve the synonym dictionary. The next repeated standardization can execute without manual review.

Figure 4– Manual review in the semi-automatic cleaning & coding tool

Evaluation data

To evaluate the validity of the proposed universal process and semi-automatic data cleaning & coding tool, 1000 diagnosis data, 1000 drug data, and 1000 examination data from the EMRs of the hospital in Beijing as the test set.

Results

The clinical data is very complex, including diagnosis, drugs, examination, and other domains. The semi-automatic tool has five text similarity algorithms to standardize clinical texts in different clinical domains. Five similarity algorithms were used in the diagnosis, drug, and examination. In Table 2, it was found that the Chinese phonetic similarity algorithm was suitable for diagnosis and drug domains, and the edit distance performed well in the examination domain. It is possible that the Jaccard coefficient, Cosine text similarity, and Tongyici cilin similarity will achieve better results in some other clinical domains such as medical devices, medical departments, etc.

Table 2– The precision of five text similarity algorithms in three different clinical domains

Text Similarity	Diagnosis	Drug	Examine
Edit distance	0.694	0.842	0.736
Jaccard coefficient	0.644	0.546	0.545
Cosine text similarity	0.632	0.825	0.358
Chinese Phonetic Similarity	0.814	0.918	0.607
Tongyici Cilin Similarity	0.305	0.444	0.260

The semi-automatic tool has no way to replace manual standardization, but it can greatly improve efficiency and reduces the time for manual work. The time of the manual coding and semi-automatic tool that handle 100 diagnosis data, 1000 drug data, and 1000 examination data were compared in Table 3. The time unit used for the experiment is seconds. The semi-automatic tool can reduce the time of manual coding by about half. The better the performance of the text similarity algorithm, the less time it takes.

Table 3– Time comparison between manual and Semi-automatic coding

		Diagnosis	Drug	Examine
Manual	Total	2233.31	1519.39	1572.63
Semi-automatic	Preprocessing	1.85	1.74	1.66
	Text similarity	6.99	8.27	7.63
	Manual review	921.35	523.20	1153.02
	Total	930.19	533.21	1162.31

Discussion

Automatic coding is important because manual coding is expensive and time-consuming. Although numerous approaches have been developed to explore automatic coding, few of them have been applied in practice[11]. This study established a universal process and a semi-automatic data cleaning & coding tool for Chinese clinical data standardization that can be used to improve the efficiency of manual coding. The tool can reduce the half time for standardization, and it was used in hospitals in Beijing. The comparison between the amount of time needed for manual and semi-automatic coding indicated the effectiveness of the tool-the time needed for semi-automatic coding takes nearly 2 times less than manual coding.

Although many studies have focused on automatic coding, we want to highlight the following advantages of our study. First, the hospital data in Beijing for research to make the tool that can be directly applied to practical work. The tool completed more than 30,000 standardizations of Chinese clinical data in

2 months, which showed high precision and efficiency. Second, the accuracy of the tool can be improved by building a high-quality synonym dictionary in the standardization process. Third, manual work could identify the shortcomings of the algorithms and strengthen the rules to improve accuracy. Fourth, the existing manual coding was improved to reduce workload and improve efficiency by our tool.

There are also shortcomings in our study. First, only semi-automation can be achieved and required manual review. Applying the advanced methods can improve the accuracy gradually, the manual review can be further reduced[22; 23]. Several studies based on machine learning approaches, such as the support vector machine (SVM)[24], natural language processing (NLP)[25; 26], deep transfer learning, etc.[27; 28], were proposed to automatic coding. The semi-automatic tool needs to combine state-of-the-art automatic coding algorithms to improve tool performance[11]. Second, it is hard to determine the appropriate text similarity algorithm in different clinical domains or even different data in the same clinical domain. Our study included the common diagnosis, drug, and examination data. Therefore, in future work, with the complete clinical domains as the goal, the semi-automation tool needs to expand the clinical domains constantly.

Conclusions

The proposed universal process was well-suited for the diagnosis, drug, and examination data standardization task and can be used gradually in other clinical domains. The proposed semi-automatic data cleaning and coding tool for Chinese clinical data standardization is feasible and practical for improving the efficiency of manual standardization and promoting clinical data quality.

Future work will include (1) clinical data standardization in more domains to test and verify the universal standardization process; and (2) improve current tool performances to close the gap between the tool and manual coding by investigating the state-of-the-art text similarity algorithms, addressing the accuracy issues in automatic coding.

Acknowledgements

This work was supported by the National Key Research and development Program of China under Grant No. 2016YFC0901703.

References

[1] J.A. Cramer, L.B. Eisenmenger, N.S. Pierson, H.S. Dhatt, and M.E. Heilbrun, Structured and templated reporting: an overview, *Applied Radiology* **43** (2014), 18.

[2] D. Gopinath, M. Agrawal, L. Murray, S. Horng, D. Karger, and D. Sontag, Fast, Structured Clinical Documentation via Contextual Autocomplete, in: *Machine Learning for Healthcare Conference*, PMLR, 2020, pp. 842-870.

[3] J. Narayanan, S. Dobrin, J. Choi, S. Rubin, A. Pham, V. Patel, R. Frigerio, D. Maurer, P. Gupta, and L. Link, Structured clinical documentation in the electronic medical record to improve quality and to support practice-based research in epilepsy, *Epilepsia* **58** (2017), 68-76.

[4] N. Hong, A. Wen, F. Shen, S. Sohn, C. Wang, H. Liu, and G. Jiang, Developing a scalable FHIR-based clinical data normalization pipeline for standardizing and integrating unstructured and structured electronic health record data, *JAMIA open* **2** (2019), 570-579.

[5] J. Kim, T.G. Macieira, S.L. Meyer, M. Ansell, R.I. Bjarnadottir, M.B. Smith, S.W. Citty, D.M. Schentrup, R.M. Nealis, and G.M. Keenan, Towards implementing SNOMED CT in nursing practice: a scoping review, *International journal of medical informatics* **134** (2020), 104035.

[6] A. Miranda-Escalada, A. Gonzalez-Agirre, J. Armengol-Estapé, and M. Krallinger, Overview of automatic clinical coding: annotations, guidelines, and solutions for non-english clinical cases at codiesp track of CLEF eHealth 2020, in: *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*, 2020.

[7] 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*, 2010, pp. 84-89.

[8] A.L. Nobles, K. Vilankar, H. Wu, and L.E. Barnes, Evaluation of data quality of multisite electronic health record data for secondary analysis, in: *2015 IEEE International Conference on Big Data (Big Data)*, IEEE, 2015, pp. 2612-2620.

[9] P. Cao, C. Yan, X. Fu, Y. Chen, K. Liu, J. Zhao, S. Liu, and W. Chong, Clinical-Coder: Assigning Interpretable ICD-10 Codes to Chinese Clinical Notes, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020, pp. 294-301.

[10] F. Teng, Z. Ma, J. Chen, M. Xiao, and L. Huang, Automatic Medical Code Assignment via Deep Learning Approach for Intelligent Healthcare, *IEEE journal of biomedical and health informatics* **24** (2020), 2506-2515.

[11] L. Zhou, C. Cheng, D. Ou, and H. Huang, Construction of a semi-automatic ICD-10 coding system, *BMC medical informatics and decision making* **20** (2020), 1-12.

[12] P. Christen and T. Churches, A probabilistic deduplication, record linkage and geocoding system, in: *Proceedings of the Australian Research Council Health Data Mining Workshop: Canberra, AU*, 2005.

[13] X. Fu and S. Ananiadou, Improving the extraction of clinical concepts from clinical records, *Proceedings of BioTxtM14* (2014), 47-53.

[14] B. Siklósi, G. Orosz, A. Novák, and G. Prószéky, Automatic structuring and correction suggestion system for Hungarian clinical records, (2012).

[15] Y. Wang, Z. Yu, Y. Jiang, Y. Liu, L. Chen, and Y. Liu, A framework and its empirical study of automatic diagnosis of traditional Chinese medicine utilizing raw free-text clinical records, *Journal of biomedical informatics* **45** (2012), 210-223.

[16] E.S. Ristad and P.N. Yianilos, Learning string-edit distance, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20** (1998), 522-532
.

[17] S. Niwattanakul, J. Singthongchai, E. Naenudorn, and S. Wanapu, Using of Jaccard coefficient for keywords similarity, in: *Proceedings of the international multiconference of engineers and computer scientists*, 2013, pp. 380-384.

[18] Y. Song and D. Roth, Unsupervised sparse vector densification for short text similarity, in: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, pp. 1275-1280.

[19] M. Li, M. Danilevsky, S. Noeman, and Y. Li, Dimsim: An accurate chinese phonetic similarity algorithm based on learned high dimensional encoding, in: *Proceedings of the 22nd Conference on Computational Natural Language Learning*, 2018, pp. 444-453.

[20] J.-l. TIAN and W. Zhao, Words similarity algorithm based on tongyici cilin in semantic web adaptive learning system [j], *Journal of Jilin University (Information Science Edition)* **6** (2010).

[21] Y. Chen, H. Lu, and L. Li, Automatic ICD-10 coding algorithm using an improved longest common subsequence based on semantic similarity, *PloS one* **12** (2017), e0173410.

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

[23] A. Rios and R. Kavuluru, Neural transfer learning for assigning diagnosis codes to EMRs, *Artificial intelligence in medicine* **96** (2019), 116-122.

[24] S. Wang, X. Li, X. Chang*, L. Yao, Q.Z. Sheng, and G. Long, Learning multiple diagnosis codes for ICU patients with local disease correlation mining, *ACM Transactions on Knowledge Discovery from Data (TKDD)* **11** (2017), 1-21

[25] R. Kaur and J.A. Ginige, Comparative analysis of algorithmic approaches for auto-coding with ICD-10-AM and ACHI, *Studies in health technology and informatics* **252** (2018), 73-79.

[26] A.N. Nguyen, D. Truran, M. Kemp, B. Koopman, D. Conlan, J. O'Dwyer, M. Zhang, S. Karimi, H. Hassanzadeh, and M.J. Lawley, Computer-assisted diagnostic coding: effectiveness of an NLP-based approach using SNOMED CT to ICD-10 mappings, in: *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, 2018, p. 807.

[27] F. Duarte, B. Martins, C.S. Pinto, and M.J. Silva, Deep neural models for ICD-10 coding of death certificates and autopsy reports in free-text, *Journal of biomedical informatics* **80** (2018), 64-77.

[28] M. Zeng, M. Li, Z. Fei, Y. Yu, Y. Pan, and J. Wang, Automatic ICD-9 coding via deep transfer learning, *Neurocomputing* **324** (2019), 43-50.

[29] Y. Yu, M. Li, L. Liu, Z. Fei, F.-X. Wu, and J. Wang, Automatic ICD code assignment of Chinese clinical notes based on multilayer attention BiRNN, *Journal of biomedical informatics* **91** (2019), 103114.

Address for correspondence

Correspondence: yanichen@zju.edu.cn (Y. Chen), lvxd@zju.edu.cn (X. Lu), tianq@zju.edu.cn (Q. Tian), chl@vico-lab.com (H. Cai).