PRACTICE GUIDELINE

Auto-Coding of Cancer Registry Data in China

Kuang Rong Wei[1], Sheng Chao Liu[2], Dongling Wei[3], Zhiheng Liang[1], Wanqing Chen[4]*

Abstract

The significance, difficulty and strategy of coding cancer data according to international coding standards are discussed, and the concept, methods and realization of cancer data automatic coding in cancer registries in China are introduced in the paper. Coding cancer data automatically with software could not only reduce the time, manpower and workload, while improving the accuracy and efficiency of cancer data coding, but also enhance the validity of cancer registration and the value of cancer registry data, which is of great significance.

Keywords: Cancer registration - data-coding - auto-coding - China

Asian Pac J Cancer Prev, **17 (6)**, 3021-3023

Introduction

Cancer registration includes such steps as collecting, sorting, coding, storing, analyzing, reporting, releasing of cancer data and so on (Jensen et al., 1991), data coding is the important step of cancer registration. Before the cancer data are used for further analysis and comparison, it should be correctively coded according to the international coding rules (Bray et al., 2014). But data coding is complex, and also very cumbersome and troublesome if manually done. Although costing lots of time and efforts, the accuracy and validity of manual data coding could not been guaranteed. Hence, developing software for coding cancer data automatically by business intelligence techniques is of important value, as discussed in this paper.

Significance of Cancer Data Coding

As some difference may exist in the definition and understanding of some cancers in different countries and areas, cancer data must be coded according to the international coding rules before being analyzed and compared. This can guarantee that what we analyze and compare are the same cancers, the bias caused by study subject can be avoided and reduced, and the research results can be more correct. Such as some difference existed in the definition and understanding of cardiac carcinoma in different areas, some carcinomas in the gastroesophageal junction, which did not originate from cardia, were mistaken as the cardiac carcinoma, and this may lead to the bias of study subjects when we analyzed cardiac carcinoma (Peng et al., 2014).

Coding cancer data according to the international coding rules is to make sure what we study are the same cancers, and to avoid the bias of study subjects. If cancer data, no matter how complete, are not coded correctly, the research result may be affected, so coding cancer data correctly is very important and the priority for scientific advance.

Difficulty of Cancer Data Coding

Cancer data coding faces many issues and problems, the main as follows: 1>. Clinical and pathological knowledge are necessary for coding cancer data correctly (IARC, 2013), but some cancer registrars may lack or short of such knowledge; 2>. International tumor especially the morphological codes are complicated, even have some issues or problems themselves. Different countries may have different understanding about the international coding rules. And the international coding rules keep update (IARC, 2015). All these make data coding difficult; 3>. Cancer registrars changed all the time, many of them were not familiar enough with the cancer data coding. 4>. Quite some registries only coded cancer data at some time such as at the end of each year or when reporting cancer data. 5>. Repeated training for cancer data coding was not so effective although its significance had been well aware of (Liang et al., 2010). Many Chinese registries existed some of above issues and problems, And it was because these issues and problems, the correctness and validity of manual cancer data coding often could not be guaranteed although much time and efforts were spent, especially for the coding of lymphoma and the tumors of central nervous system.

[1]*Zhongshan Cancer Registry, Zhongshan,* [4]*Chinese National Office for Cancer Prevention and Control, Beijing, China,* [2]*Department of Technology, Dimensional Insight, Massachusetts,* [3]*Computer Science Department, Dartmouth College, New Hampshire, USA*
**For correspondence: mrchenwq@yahoo.com*

Strategies for Simplifying Cancer Data Coding

Some methods were used by related experts and institutions worldwide to simplify and facilitate cancer data coding, such as arranging the international diseases codes in the order of alphabet, spelling and stroke, and these ways did easy data coding and improve its accuracy and validity to some extent. The book about international diseases codes, compiled by the Organization of International Code for Diseases (ICD), included the chapters about searching the disease codes by the order of English alphabet (IARC, 2013; 2015), and had been translated into different languages to be used in many countries and areas, such as Chinese (National Office for Cancer Control and Prevention et al., 2004), to make it more convenient to be used. Shanghai Municipal Center for Disease Control and Prevention compiled the book Tumor Nomenclature and Coding, which introduced the methods and skills for cancer data coding, such as comparing the codes of ICD-10 and ICD-O-3, putting them in the order of Chinese spelling, and providing quick ways for searching the codes of common English abbreviation of cancer names (Lu et al., 2011). Currently, the codes of ICD-10 and ICD-O-3 can be fuzzily searched in most cancer data management software. When you put cancer names into these software, the software will show you some corresponding codes for you to choose, and vice versa. This could simplify data coding and improve the accuracy and validity of data coding.

CanReg5, the software for cancer data management, provided by International Agency for Research on Cancer (IARC), have the same function of fuzzily searching ICD codes too (Morten Johannes Ervik, 2014). Other software provided by IARC could convert ICD-10 and ICD-O-3 codes mutually (J. Ferlay et al., 2005). The website of Chinese cancer control and prevention database (http://cancernet.cicams.ac.cn) already provided the methods for searching ICD-10 and ICD-O-3 codes at an earlier time, and some companies provided the software for facilitating ICD-10 and ICD-O-3 coding (http://www.hdcsc.com/). In 2013, Wang qingsheng et al developed a smaller and more convenient ICD Fuzzy query software, which could improve the efficiency and validity of cancer data coding (Wang et al., 2013).

Coding Cancer Data Automatically

Although above methods could somehow improve data coding velocity and validity, it was still very troublesome and cumbersome. Sometime even often, coders just choose the codes they thought were right without confirming further when they were not sure which codes was right. So we wonder if a more friendly software, which can code cancer data automatically, can be developed on the basis of previous work. If yes, cancer registrars can be liberated from the tedious and cumbersome coding work at large extent.

Skilled coders could code cancer data quickly and accurately. If the principles and rules of ICD coding, the methods, techniques and thoughts of the skilled coders can be reorganized in such ways that they can be expressed by computer languages, then a software, which can code cancer data automatically, may be developed and replace most manual coding.

Based on above ideas, before developing the cancer data automatic coding software, we reviewed systematically the principles and rules of ICD coding, and thoroughly explored the coding methods, techniques and ideas of the skilled coders. ICD includes two main parts which are anatomic and morphological codes. Anatomic codes are determined by the sites and subsites the tumors originates, and morphological codes are determined by the tissues and cells of the tumors occurs, the levels of tumor differentiation, and the behaviors of tumor biology. When we code the tumor anatomic codes, the bodily systems where the tumors occurs, such as digestive, respiratory, urinary, genital, hematological and central nervous system etc, should be identified first, then the organs the tumors originates, such as esophagus, stomach and colon in the digestive system, be identified, and finally the subsites of organs the tumors occurs, such as cardia, pylorus, lesser and greater curvature of stomach, should be identified.

When developing the software, the same processes or principles were followed too. It meant that when coding tumor anatomical codes automatically by the software, first the systems, organs and subsites of the organs the tumors occurred should be in turn identified, then the anatomical codes of ICD-10 and ICD-O-3 be automatically assigned by the software accordingly. Coding morphological codes are more complicated. First, the tumor should be decided if or not should be registered, this was determined by the biological behavior of the tumor. The following tumors should be registered: all malignant tumors(the behavior codes are 3, 6 and 9), some in situ cancers (the behavior code is 2), benign (the behavior code is 0) and uncertain behavior tumor (the behavior code is 1) of central nervous system. Secondly, the levels of tumor differentiation, classified as undifferentiated, low-, middle- and high differentiated, should be identified. Thirdly, the types of tissues which the tumor originated, such as epithelium, mesenchyme, lymph, hematopoietic and nerve, should be identified. Finally, the specific cells which the tumor originated, such as squamous, gland, basal and transitional cells in the epithelial tissues, should be identified. When developing the software, the same processes or principles were followed. It meant that when coding tumor morphological codes automatically by the software, the software first judged if or not the tumor should be registered according to the international coding rules, if yes, the software in turn identified the differential levels of the tumor, the specific tissues and cells the tumors originated, and finally assigned automatically the morphological codes accordingly.

The above principles were the general principles for developing the software, some special situations existed and should be considered carefully, such as the English character C and D are used to distinguish the tumor biological behaviors by the ICD-10 coding rules, and the anatomic codes of all types of leukemia are C42.1 by the ICD-O-3 coding rules (National Office for Cancer Control and Prevention et al., 2004; J. Ferlay et al., 2005; Lu W et al., 2011; Morten Johannes Ervik, 2014). The same disease, such as Hodgkin disease, may be expressed in different Chinese characters in different areas and by different persons. And some diseases were named after someones names. All these increased the complexity of data coding.

Testing of the Software

The software, developed by the intelligent business techniques and according to the above principles, was strictly tested before being used routinely. First, the testing was carried out in Zhongshan cancer registry. The testing method was to code the same cancer data by the software and manual ways simultaneously, and compare the fitness between them. If the coding results were not consistent, the results would be reviewed carefully. If the results of software coding were confirmed wrong, the software would be adjusted and modified according until totally right. Often the discrepancy between them was caused by the wrong manual coding, so, the software could be used to check if or not the manual coding were wrong. When the fitness between them was over 95 percent, the software was tested in many other cancer registries such as the ones in Canton, or with the data from the registries in Hebei and Yunnan provinces of China. The testing results also confirmed that the software could code cancer data accurately, with the accuracy rate more than 95 percent, if the tumor diagnoses were expressed or written properly, although some problems and issues, mainly caused by the cancers names unseen before or expressed differently, may existed. For more convenient and friendly to be used, the software was also modified so that it could identify the different expressing ways of cancer name in different areas, such as the different expressions of Hodgkin disease in Chinese character.

Advantages and Disadvantages of the Software

More than 2 years practical application had shown that the software had the following advantages:1> Running quickly. 10,000 cancer cases could be coded automatically in only about 10 minutes, but it may need 2 months for a skilled coder to finish the same work. 2> Accuracy. The accurate rate of the software was more than 95 percents, higher than the rates of most skilled coders. Manual coding was affected by individual factors such as emotion and seriousness, but coding by software was not. The mistakes made by the software were mainly caused by some unseen, rare-seen or un-proper expressed tumor names.3> Greatly reducing the tediousness and cumbersomeness of manual data coding, and saving lots of manpowe.4> Could be used to check the correctness of manual data coding.

The disadvantages of the software were as follows: 1> The accuracy and validity of the software would be greatly affected if the tumor names were not expressed properly or accurately,. 2> When the ICD rules are changed or updated, the software would need to be modified accordingly.

Significance of Auto Data Coding

Technology always changed our life and works, just as it changed medical records and practices. auto cancer data coding could change cancer data coding too. It could free data coders from tediousness and cumbersomeness data coding, improve the speed and accuracy of data coding, enhance data quality, guarantee the correction of related study results, facilitate and promote international exchange, except for probably reducing the necessary of data coding training. The most imprtant is that cancer registrars can do more other meaningful works such as data analysis after being liberated from manual data coding.

In general speaking, auto cancer data coding by software was not only quick and accurate, but also represented and embodied the developing trend of artificial intelligence ,and can be used in other medical and health fields according

References

Bray F, Znaor A, Cueva P, et al (2014). Planning and developing population-based cancer registration in low-and middle-incoming setting. Lyon: IARC Technical publication NO. **43**, 5-18.

Ervik MJ (2014). CanReg5. Lyon:IARC, **49**.

Ferlay J, Burkhard C, Whelan S, et al (2005). Check and conversion programs for cancer registries (IARC//IACR Tools for Cancer Registries). Lyon: IARC Technical Report No. **42**, 4.

IARC (2013). International classification of diseases for Oncology, 3rd Edition, First Revision[ON/DB].Geneva: World Health Organization.

IARC (2015). International statistical classification of diseases and related health problems, 10th Revision. [ON/DB]. Geneva: World Health Organization.

Jensen OM, Parkin DM, MacLennan R, et al (1991). Cancer registration: principles and methods. Lyon:IARC Scientific Publication No. **95**,7-21.

Liang ZH, Liu J, Wei KR (2010). Some views about cancer registration. *China Cancer*, **12**, 779-781.

Lu W, Zheng Y (2011). Tumor nomenclature and coding. Shanghai. Press of the Second Military Medical University, 3-6, 116-209, 250 [in Chinese].

National Office for Cancer Control and Prevention, Information Center for Health Statistics of Health Ministry, National Center for Cancer Registration (2004). Guideline for chinese cancer registration. Beijing: Press of Chinese Union Medical University, 50-8 [in Chinese].

Peng XB, Chen WQ, Chen ZF, et al (2014). Cardiac cancer epidemiology in China. *Chinese Arch General Surg*, (Electrical Version), **8**, 61-4 (in Chinese).

Wang QS, Chen WQ (2013). Fuzzy query system for cancer data coding of ICD-O-3 and ICD-10. *China Cancer*, **5**, 369.