Biotech Health Sci. 2018 February; 5(1):e13667.

Published online 2018 February 1. **Research Article**

Designing Automatic Coding Module of Cancer Open Text Pathology Reports Based on International Classification of Diseases for Oncology

Peyman Rezaei-hachesu[1,2], Nazila Moftian[1], Mahsa Dehghani[1], Taha Samad-Soltani[2*]

[1]Department of Health Information technology, Tabriz University of Medical Sciences, Tabriz, Iran
[2]Road Traffic Injury Research Center, Tabriz, Iran.

*Corresponding author: Taha Samad-Soltani, School of Medical Informatics and Management, Golgasht Street, Tabriz, Iran. Tel: +98 912 9321 546. Email: t-ssoltany@razi.tums.ac.ir.

Received 2017 March 12; **Accepted** 2017 June 14.

Abstract

Background: In the domain of clinical documents, all diseases are classified at templates by the world health organization and specific codes have been assigned to them. The goal of this study was automatic coding of cancer free texts based on International Classification of Diseases for Oncology (ICD-O-3) and evaluation of results.
Methods: In this research, the preparation and development of one initial sample of automatic coding module on pathology reports open texts existing in PubMed's cancer titles database is performed for exploitation of the information based on the texts related to cancer to coding the information based on ICD-O-3. After developing the algorithm for exploiting cancer phrases and the codes based on ICD-O-3 and converting them to code in programming environment, the required data for implementation and algorithm testing were performed and finally the obtained results were evaluated.
Results: Automatic coding prepares the possibility of coding and listing information inside the text and also coding the existing titles of neoplasms at descriptive text of pathology reports and with an accuracy of approximately 70%. This study explained a simple stepwise approach to coding issues in medicine.
Conclusions: It performed effectively on free texts and could be used as a decision support module in Health Information Systems to reduce coding errors.

Keywords: Clinical Coding, Informatics, Neoplasms, Pathology.

1. Introduction

Recognition and detection of text mining, which aims at separating the knowledge of text data is a critical operation(1, 2). This approach is mostly used in specialized biol medical areas where language templates are used broadly for text based documents (3). The widespread use of medical controlled vocabularies is appeared in response to the need for data exchange and standardization at research and modern medical care. The most popular method used for such coding is human coding, which is utilized in both medicine and classification system disciplines(4, 5)

Clinical records can include beneficiary information in form of open texts (6). In electronic medical records, the information such as family history, symptoms, signs and smoking history are usually included descriptively in text fields that are written by clinical specialists as progress notes or as a discharge summary. When the coded data (including insurance codes for recognitions and deaths) are prepared, they are not always correctly coded or may be manipulated because of credit or financial reasons(7).

Medical language extraction and encoding systems (MeDLEE) were introduced first time(8). These systems as a natural language processing (NLP) are used to encode medical records in a structured format. The spread of natural language together with consistent environment of health care have abled the staff to describe clinical information. This technique has become a popular method to exploit information from open texts using natural language processing and also promises the possibility of presenting useful medical information from a large volume of data automatically and for various goals with lowest human efforts. The initial setup and cost will require some time and effort, but the solution will save and increase health if was used in an evidence based manner (9, 10).

Identification of phrase structure in natural language is performed using a finite state machine which is a translated form of one regular expression(11, 12). A regular exn pression is a sequence of characters which helps to find or adapt the string or a set of strings using specific grammar which is held as a pattern (13). This method is used in a large number of areas.

In the domain of clinical documents, all diseases are classified at templates by the world health organization and specific codes have been assigned to them. International classification of diseases for Oncology (ICD-O) is basically applicable at tumors or cancer registries to encode neoplasm form and location which these texts are

usually obtained from pathology reports(14). Software products which perform calibration and encoding of medical text are called autocoding(15).

In this research, development of a basic autocoding module is performed for extraction of the information from the texts related to cancer due to encoding the information based on ICD-O. This study performs on open text of pathology which exists at data base of PubMed cancer titles. A number of these reports are selected randomly and will be used for module evaluation.

2. Method

The approach and method of this research similar to previous studies requires three steps which are:

1. Development of an algorithm and a definition for the intended concept due to extraction (cancer terminology and the codes based on ICD-O),
2. Preparing data (pathology report) to implement and test algorithm and data processing for code extraction and finally
3. Results evaluation. The steps of implementation are addressed in following.

2.1. Development of algorithm and extraction of concept based on ICD-O codes

Computational algorithm due to implementing autocoding has eight general steps shown as flowcharts in Fign ure 1. These include the step of allocating cancer words to demonstration and registration of output.

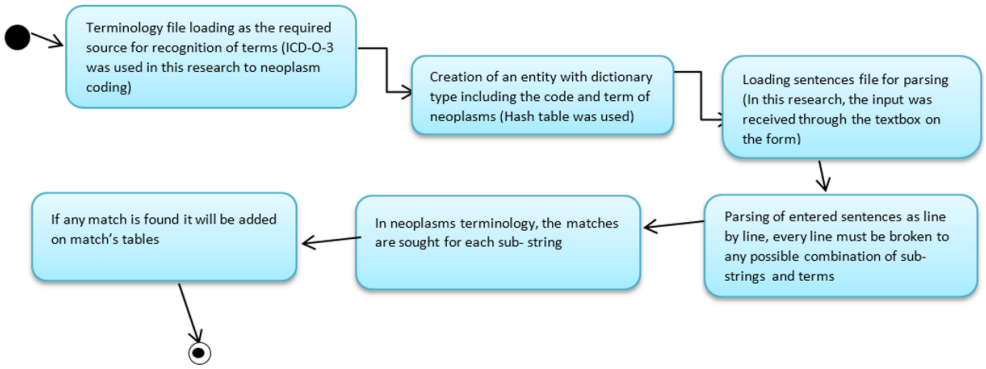

Figure 1. The suggested algorithm for autocoding of neoplasm

It is necessary to classify and organize the dictionary included with all terms and ICD-O codes to extract the terms from text. For this purpose, the text file of data set including the ICD-O-3 codes and terms of national cancer institute of the United State are used(16). The template of this file is as code and the phrases are in front of it which one sample of that is shown in code segment 1 .

First, a regular expression was needed to recognize, extract and locate them individually within an appropriate data structure to apply these terms and codes at programming environment which this regular expression was written as a code segment in code segment 2.

Analysis of this code is in a way that each recognized expression has to include two elements of code and term and code structure is in the form (?<Code>\\b\\d{4}/\\d(1)) which means that the recognized expression includes a part named code. \b means position match in the beginning or in the end of each word. \d represents digits and numbers which are 4 digits ({4} demonstrates that digits number is consecutive) and after, the '/' character comes and then one number appears again. Using // in place of / is because of programming compiler translation style.

The expression (?<Term>\\s(1)[a-zA-Z]+[, a-zA-Z]* demonstrates the existing of a second component with term title. \s matches each whitespace character. Therefore if there is one space (with number 1 (1)) at the beginning of disease titles, it will not be considered. Then one or more characters comes after which can be little or capital letters from 'a' to 'z'. The character '+' demonstrates the existing of at least one character and at most with desired characters. After this set of characters which constitute disease or injury title, a combination with the number of 0 or more (* means zero or more) as a character of ',' following by blank space and again a combination of English letters.

Code lines of RegexOptions.IgnoreCase and RegexOptions.IgnorePatternWhitespa causes not considering letters capital state and whitespaces in the beginning and the end of discovered templates.

Above regular expression has the ability of recognizing codes and the expressions related to each code as two-components. Two discovered components are inserted within hash table to have an easy and simple access, search and reference to these concepts.

In the next step, searching matched terms and allocating code to them must be performed on the desired text related to the oncology. Before matching on the desired entered text, for consistency of text, the words "tumor" and "oma" including uncommon forms become converted to a similar template which it has been performed in the code of code segment 3.

```
Carcinoma, NOS 8010/3

Epithelioma, malignant 8011/3

Large cell carcinoma, NOS 8012/3

Large cell neuroendocrine carcinoma 8013/3

Large cell carcinoma with rhabdoid pheno- 8014/3
type

Glassy cell carcinoma 8015/3
```

Code segment 1. A part of data set structure of ICD-O implications presented by United States cancer national institute

```
public static Regex MyRegex = new Regex(
            "(?<Code>\\b \\d{4}/\\d(1))(?<Term>\\s(1)[a-zA-Z]+[, a-zA-Z]*" +
            ")\r\n\r\n\r\n\r\n\r\n",
            RegexOptions.IgnoreCase
          | RegexOptions.IgnorePatternWhitespace
          );
```

Code segment 2. Regular expression to recognize codes and the terms in open texts.

```
public static Regex singular = new Regex("omas",RegexOptions.IgnoreCase |
RegexOptions.IgnorePatternWhitespace);

        public static Regex Engular = new Regex("tumo[u]?rs", RegexOptions.IgnoreCase
| RegexOptions.IgnorePatternWhitespace);
...
...
sen= singular.Replace("oma", sen);

sen = Engular.Replace("tumor",sen);
```

Code segment 3. A part of conversion code and equalization of multi-form words from an un-common form to a common

In the next step, the algorithm tries to matches the contents within the hash table for each row with separation of arbitrary text words from space lines and pouring it inside a string. In this step, all combinations of separated words of each row is compared with existing expressions in hash table and in the case of adaptation will be assigned to that code row or that key of hash table. We used *Permutations generator source code* to generate all combinations in a single row(17).

When this code splits the string to some words, all permutations of these words were generated and combined, for example: if A, B and C are split, then algorithms generated ABC(A+'[SPACE]'+B+'[SPACE]'+C), ACB, BCA, CBA, BAC, CAB, ABC, AB, BC, AC, CA, A, B, C. these combinations compared with hash table values (from biggest combination to smallest). The algorithm was stopped when exact or semi exact result was obtained. The code of resulted value added to data grid.

3. Results and Discussion

3.1 Preparing data for implementation and testing algorithm

In this step to test algorithm, the set of cancer titles prepared from PubMed was used which includes over 18000 scientific texts with cancer theme and as row by row(15). Above template is performed on 100 titles of this file which is selected randomly and is implemented as batch and each row is encoded separately. A sample of results on the interface of written program is shown in Figure 2.

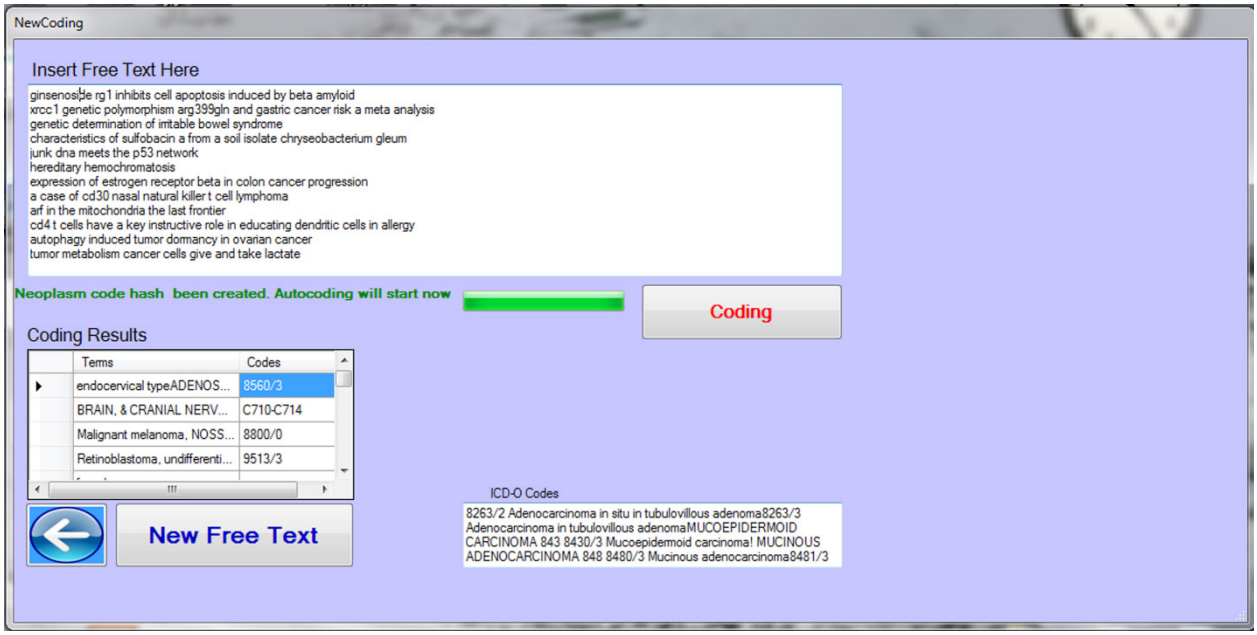

Figure 2. Program interface and a sample of coding results on the set of cancer title

Assigned codes by the designed module were compared with the coding of experts and evaluation of outputs was performed based on sensitivity, accuracy and precision analyses. To measure sensitivity, accuracy and specificity, for equations 1, 2 and 3 were used.

3.2 Results evaluation.

Sensitivity is the ratio of correct positive cases which is correctly declared by a certain test, which based on this definition, 100% sensitivity means the correct recognition of all patients(18).

$$Sensitivity = \frac{Number\ of\ correct\ positive\ recognitions}{Number\ of\ correct\ positive\ recognition + Number\ of\ wrong\ negative\ recognitions} \quad (1)$$

Statistical scale determining test precision for recognition of negative cases is called specificity.

$$Specificity = \frac{Number\ of\ correct\ positive\ recognitions}{Number\ of\ correct\ negative\ recognition + Number\ of\ wrong\ negative\ recognitions} \quad (2)$$

And finally to calculate accuracy the following formula is used:

$$Accuracy = \frac{Number\ of\ correct\ positive\ recognitions + Number\ of\ wrong\ negative\ recognitions}{Number\ of\ correct\ negative\ recognition + Number\ of\ wrong\ positive\ recognitions + Number\ of\ correct\ positive\ recognitions + Number\ of\ wrong\ negative\ rexognitions} \quad (3)$$

The obtained results on the analysis of 100 sentences of cancer titles are shown in confusion matrix in Table 1.

Table 1. Confusion matrix of evaluating automatic encoding application

		Computer	
		Wrong	Correct
Expert	Wrong	0	0
	Correct	31	69

The obtained results are shown by system with the encoding results of medical documents expert in this table. Because the criteria of comparing automatic encoding result was with expert encoding with the supposition of 100% accuracy of his coding, thus the values in the first row became zero. From 100 sentences entered to the system, 69 was correctly coded and was added to output table and 31 cases were coded wrong or unidentified and had defect or inconsistency.

Evaluating sensitivity, precision and accuracy with exploitation of relations 1, 2 and 3 are equal to 100, 0 and 69 percent respectively which the third criteria i.e. system's general accuracy is the best criteria at correct or wrong conditions.

4. Discussion

Computer documents include useful information but finding the thing that people desire is challenging and complex because that information can only be found in text. Autocoding enables information to be listed and encoded in text. It involves a natural language parser, an inference engine and Lexicon list. It also, along with a special vocabulary searches the concept in the text(19). For example, in this research exploiting lexicon list ICD-O-3, autocoding could find the neoplasm titles in descriptive texts of pathology reports and encodes them. Of course it has to be mentioned that autocoding of cancer pathology reports is only one of the applications.

Another research which was performed in Ontario cancer registration system, because of existing problems at pathology reporting process with paper and its high information and material costs including reports misses, delay at result announcement, lack of sending paper results by laboratory and report retrieve costs was used among tens of thousands previous reports from pathology report electronic system. In a part of this project, autocoding technology was used based on ICD-O-3 and SNOMED and it was shown that exploiting the combination of these technologies, the costs of each report was reduced from 3.35 $ to 1.63 $(20, 21). In this research bea cause of lacking operational implementation, cost-effectiveness analysis is not performed. With regards to many benefits of cancer pathology experiments reporting computerization it is suggested that designing cancer registration systems be performed with reporting approach on line and autocoding methods be performed locally and as much as possible on physician writings.

To evaluate the methods based on natural language it has been suggested that instead of algorithm evaluation, outputs efficiency, accuracy and precision be evaluated. In applying this technology, its applications must be known before evaluation; how to exploit it within organization procedure must be closely recognized; current performance of human forces and their efficiency has to be evaluated in comparison with the new method of investigation critically and the methods of updating, promoting and improving this technology must be considered at organization procedure(22). In this research, accuracy analysis method based on system output was used. But because of the mentioned limits around implementation and integration of this system at organization process, implementing comparative evaluations were not prepared. Some other intelligent algorithms were proposed to optimize performance of solution (23). We suggest the application of such algorithms in improvement of computerized medical coding.

In 2016, Wei et al proposed an autocoding software for coding cancer registry in china. They achieved 95% accuracy with a manual confirmation step (24). In an other study, Pakhomov et al, reach a 82% accuracy in an auto assignment of Diagnosis Codes to Patient Encounters module. They used machine learning algorithms (25). Cernile mentioned in a report entitled "Automated Classification of Cancer Pathology Reports", that an accuracy more than 70% is a good outcome, but we need more than 95% to be effective (26). In current study we achieved 69% acw curacy without any manual confirmation. The achieved accuracy satisfied minimum requirements. if confirmation methods had been used in next steps, accuracy increased dramatically.

4.1. Conclusion

This research presented a system having simple and brief code which registered promising results. It can be suggested that the presented approach can be used to promote performance and also applying it in more common areas including assurance encoding, diseases and drugs. In the field of biomedical informatics, it is necessary to exploit medical terms from text and adding them to a specific lexicon list. To improve accuracy, manual confirmation steps are necessary. Future studies will focus on big data handling by improving and optimizing of knowledge extraction algorithms.

Ethical consideration: Not Applicable

Implication for health policy makers/practice/research/medical education: Automatic detection of diseases on open text medical reports can facilitate processing of large volume of data and improve documentation as well as error reduction.

References

1. Meloni V, Sulis A, Ghironi D, Cabras F, Del Rio M, Monni S, et al. HL7apy: a Python library to parse, create and handle HL7 v2. x messages. EJBI. 2015;11(2).
2. Sheshasaayee A, Jayanthi R. A Text Mining Approach to Extract Opinions from Unstructured Text2015.
3. Vincze V, Szarvas G, Farkas R, Móra G, Csirik J. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. BMC bioinformatics. 2008;9(Suppl 11):S9.
4. Kornai A, Stone L. Automatic translation to controlled medical vocabularies. STUDIES IN FUZZINESS AND SOFT COMPUTING. 2004;140:413-34.
5. Rezaei-Hachesu P, Samad-Soltani T, Khara R, Gheibi M, Moftian N. 192: PREDICTION OF ASTHMA CONTROL LEVELS USING DATA MINING METHODS: AN EVIDENCE-BASED APPROACH. BMJ Open. 2017;7(Suppl 1):bmjopen-2016-015415.192.
6. Priya R, Padmajavalli R. Biomedical Text Mining for Diagnosing Diseases - A Review2016.
7. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC medical informatics and decision making. 2006;6(1):30.
8. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. Journal of the American Medical Informatics Association. 2004;11(5):392-402.
9. Stephen R. Context identification in electronic medical records: Massachusetts Institute of Technology; 2004.
10. Peyman RH, Ahmadi M, Aziz R, Zahra S, Farahnaz S, Nader M. Clinical care improvement with use of health information technology focusing on evidence based medicine. Healthcare informatics research. 2012;18(3):164-70.
11. Evans DK, Klavans JL, Wacholder N, editors. Document Processing with LinkIT. RIAO; 2000.
12. Janani R, Vijayarani S. An Efficient Text Pattern Matching Algorithm for Retrieving Information from Desktop. Indian Journal of Science and Technology; Volume 9, Issue 43, November 2016. 2016.
13. Python regular expressions: Copyright © tutorialspoint.com; 2013 [Available from: http://www.tutorialspoint.com/python/python_reg_expressions.htm.
14. Mendonça EA, Lussier YA. The Frontiers of Computational Phenomics in Cancer Research. An Omics Perspective on Cancer Research: Springer; 2010. p. 201-10.
15. Berman JJ. Methods in Medical Informatics: Fundamentals of Healthcare Programming in Perl, Python, and Ruby: Taylor & Francis; 2010.
16. Berman JJ. ICDO3 txt file Surveillance, Epidemiology and End Results (SEER) Program: National Cancer Institue; 2010 [Available from: http://seer.cancer.gov/data.
17. Treenef. Permutations generator source code 2005 [Available from: http://www.cprogramming.com/snippets/source-code/permutation-generator.
18. Samad-Soltani T, Ghanei M, Langarizadeh M. Development of a Fuzzy Decision Support System to Determine the Severity of Obstructive Pulmonary in Chemical Injured Victims. Acta Informatica Medica. 2015;23(3):138.
19. AUTOCODE Extracting Meaning from Text2013. Available from: http://www.aim.ca/pdf/AutoCode.pdf.
20. Dale D, Golabek J, Chong N. The impact of E-path technology on Ontario Cancer Registry operations. J Registry Manage. 2002;29(2):52-6.
21. Johanna Johnsi Rani G, Gladis D, Mammen JJ, Manipadam MT. Tumour Classification and Analysis from Breast Cancer Pathology Reports using Natural Language Processing2015.
22. Wolniewicz R. Computer-assisted coding and natural language processing. 2015.
23. Nabaei A, Hamian M, Parsaei MR, Safdari R, Samad-Soltani T, Zarrabi H, et al. Topologies and performance of intelligent algorithms: a comprehensive review. Artificial Intelligence Review. 2016:1-25.
24. Wei K, Liu S, Wei D, Liang Z, Chen W. Auto-Coding of Cancer Registry Data in China. Asian Pacific journal of cancer prevention: APJCP. 2016;17(6):3021.
25. Pakhomov SV, Buntrock JD, Chute CG. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. Journal of the American Medical Informatics Association. 2006;13(5):516-25.
26. Cernile G. Automated Classification of Cancer Pathology Reports. NAACCR 2011 Conference: 2011.