MEDINFO 98
B. Cesnik et al. (Eds)
Amsterdam: IOS Press

Integration of the analytical and alphabetical ICD10 in a coding help system. Proposal of a theoretical model for the ICD representation

Christine Bouchet[a], Olivier Bodenreider[a], François Kohler[a]

[a]*Laboratoire SPI-EAO - Faculté de Médecine de Nancy - BP 184 - 54505 Vandoeuvre Cedex - France*

Abstract

In French hospitals, medical diagnosis coding with the ICD10 is commonly performed and the use of effective tools would help coders in their task.

Aim of this work: to ameliorate an existing coding help system. This system, which already consists of the ICD10 analytical index, would be increased with the terms of the alphabetical index that includes lexical variants and additional terms as well. The addition of the second volume of the ICD would allow the coding of more terms and would lessen documentary silence.

Methods: the first step of this work was a careful study of a theoretical model of the ICD content. Then the alphabetical index file was submitted to a lexical analysis, and it was automatically transformed to be integrated into the existing coding help system.

Results: Compromise had to be made between a theoretical model and between what could be obtained in practice by an automatic processing of the file. Finally the alphabetical index was added to the initial thesaurus, which represents 42,000 terms and 4,000 additional words. Links between words and codes were also considerably increased, which has enhanced the searching possibilities of the tool and lessen documentary silence. Conversely the research time has been increased.

Conclusion: difficulties have to be encountered when trying to turn a manual tool into an automatic research tool.

Keywords

Coding; Disease Classification; Medical Diagnosis; ICD10; Knowledge Representation

Introduction

Coders need effective tools that would help them in their daily medical diagnosis coding. In France most of the time this coding is not performed by medical recorders but by people who are not very familiar with the three volumes of the ICD10 (International Classification of Diseases Tenth Revision) [1]. The development of computerized coding help systems could allow coders to have access to all the information available in the ICD and to retrieve them rapidly.

Up to now, these tools are far from being satisfactory and while carrying their research, users have to cope with documentary noise (too many non relevant results) or with documentary silence (no results)[2]. Therefore, these tools have to be urgently improved.

Three main types of coding help systems can be described.

Lexical coding help systems

They enable a literal string research in a sentence. They can easily be implemented but they often result in documentary silence. For example, should the term "febris rubra" be searched in the ICD, since the terms used in the analytical volume are "scarlett fever" and "scarlatina", no answer would be given. The main interest of these tools is to enable people who are not familiar with the ICD to code a term without knowing the hierarchical organization of the classification.

Tree browsing and hypertext

These systems are hierarchically organized like the initial classification and every level allows a visualization of the lower levels. To be used, these systems require a good knowledge of their structure.

Documentary systems

In this case, coding is carried out with the help of a thesaurus, which allows defining lexical variants and linking alternative views of the same concept together. It also identifies relations between different concepts as in the semantic network included in the UMLS (Unified Medical Language System) [3,4] or in the MED (Medical Entities Dictionary) [5]. With such systems a research on the string "heart" would allow the coding of "mitral valve prolapse", since the mitral valve is a part of the heart.

Such systems have to be rapid and easy to use. They also have to offer a program and table contents both completely independent, so that the updating could be quickly achieved and the researches could be carried out with the help of several classifications or nomenclatures [6]. Some systematized nomenclatures like SNOMED, used in restricted area, enable a completely automated coding [7]. In each case, automated systems have to be based on elaborate models of knowledge representation [8,9,10].

A coding help system has been used in our department for several years [11]. It belongs to the documentary type and can partially perform synonym research. It has been fed with information contained in the analytical volume of the ICD10; cross-references, exclusions and comments being kept in its structure. The aim of our work was to add terms coming from the alphabetical index to the existing application in order to lessen documentary silence. First, careful thought has been given to the structure of the alphabetical index. Then a computerized process has been used to be included into the coding help system.

Material and methods

Existing coding tool

The initial application is a coding help on microcomputer implemented with the relational database software Fourth Dimension. This coding help system includes the French version of the ICD9, ICD10 and of the CDAM (Catalogue des Actes Médicaux). The medical diagnosis coding part consists of:

- an ICD code table including a code, its term and the corresponding inclusions and exclusions.
- a thesaurus with all the words existing in the terms and an identifier for each word.
- a table allowing the implementation of the links between the words of the thesaurus and the codes.

In the application, research can be carried out:

- either by codes
- either by words. For example a research with the word "infarction" will bring out all the records including the string "infarction". A research with the words "myocardial infarction" will bring out the records including both the strings "myocardial" AND "infarction".
- Lexical variants and synonyms are partially treated. e.g. ischaemia and ischaemic (lexical variants) or gastric and stomach (synonyms) will have the same identifier. So a research with the words "stomach disorder" will bring out the records corresponding to "gastric disorder".

However, this database only includes lexical terms coming from the analytical index and not those from the alphabetical one. This alphabetical index encompasses more numerous terms than the analytical one. Additional terms can either be lexical variants of strings from the analytical index whereas completely new terms. With the initial coding help system some researches cannot be carried out. E.g. the Arthus' phenomenon which could not be coded with the initial coding help tool since in the analytical index the corresponding main term is "Allergy, unspecified". Only a tool including the two indexes enables this type of research.

Theoretical model

To take into account the alphabetical ICD10 index, a careful thought has been given to the modeling of the data representation. In the initial application the model was simple. Beside the traditional hierarchy in codes, subcategories, categories and chapters, each main term had a code (relation 1-1). The inclusions were text fields linked to the codes. The adding of strings coming from the alphabetical index has brought new strings for a given code. Then the question was how to include these new strings among the already existing ones.

Alphabetical ICD10 index processing

This alphabetical index had to be transformed in order to be added to the existing database. It consisted of a text file corresponding to the alphabetical volume of the French version of the classification. The file included 50,000 strings.

First a lexical analysis of this file was performed in order to identify all the components of each single line. Then the file was automatically rewritten with tags inserted between each component of a line in order to facilitate the inclusion in the database.

Integration of the alphabetical index into the database

The resulting file has been integrated into the coding help system. The added strings were separated into words, which enabled to add records to the key-words table and links between codes and words.

Results

Theoretical modeling approach

In a conceptual representation ICD can be seen as being composed of four categories of objects hierarchically displayed codes, subcategories, categories and chapters. Properties can be linked to each of these objects.

An additional hierarchical level can be added. Inclusions or additional terms coming from the alphabetical index are either synonymous strings of main terms or real included strings. The synonymous strings are on the same hierarchical level as the main terms while included strings should be on a lower level. But they cannot be automatically differentiated in the ICD. E.g.:

I20.9 Angina pectoris, unspecified

Anginal syndrome

Ischaemic chest pain

In this case "anginal syndrome", "ischaemic chest pain " and "angina pectoris" are synonymous.

I25.3 Aneurysm of heart

Aneurysm

- mural
- ventricular

In this case "aneurysm mural" and " aneurysm ventricular" are precisions for aneurysm, they should be on a lower level.

The final result is a series of objects fitted together in a hierarchical way and owing properties such as inclusions or exclusions. Each level inherits the properties of the upper level except when the inheritance is blocked by the redefinition of a property at the lower level. Anyway in some cases the upper level property does not make any sense at the lower level. e.g.

I21: Acute myocardial infarction

Includes: myocardial infarction specified as acute or with a stated duration of 4 weeks or less from onset.

I210 Acute transmural myocardial infarction of anterior wall

I211 Acute transmural myocardial infarction of anterior wall.

In this case it is a real inheritance. The property (inclusion) is relevant for each lower level.

Q45: Other congenital malformations of digestive system

Excludes: congenital:

diaphragmatic hernia (Q79.0)
hiatus hernia (Q40.1)

Q45.1: Annular pancreas.

Here the exclusions are not relevant for the lower level.

Alphabetical ICD10 index processing

The lexical analysis of the components of the alphabetical list has shown that each line is composed of:

- an indented string
- supplementary words in parentheses (optional)- cross-references (see or see also) (optional)
- one or two codes (except in the case of the cross-reference "see")

There is no tag separating the different components. e.g. :

Hypophosphatemia (acquired) (congenital) (familial) (renal) E83.3

Hypopiesis - see Hypotension

Hypopharyngitis (see also Laryngopharyngitis) J06.0

Some lines can have several codes for one string (frequent with dagger-asterisk codes). e.g.:

Pericarditis in systemic lupus erythematosus M32.1+I32.8*

The whole file has been automatically restructured. One line has been written for each code and a tag has been inserted between the different components (code, string, supplementary words, and cross-references). Then the data have been written in the resulting file according to the following order: a code - a tag - a string with supplementary words in parentheses. e.g.:

M32.1+ [tabulation] Pericarditis in systemic lupus erythematosus

I32.8* [tabulation] Pericarditis in systemic lupus erythematosus

Cross-references have been listed in separate files.

Integration of the alphabetical index in the database

In theory, strings coming from inclusion fields or from the alphabetical index could be divided into two categories those, which are equivalents of the main terms and those. But in practice such a division is not automatic. The only solution would have been to treat each string individually, which did not correspond to an easily reproducible processing. The model was therefore simplified and all the strings have been implemented in the same way using a 1-n relation (1 code for n strings).

The initial model has been modified. The inclusion text fields have been broken up into strings and a string table has been created; it contains the main terms and their inclusions. The new strings coming from the alphabetical index have been added and cut into words, which brought new words into the thesaurus and new links between the words and the codes.

The initial text file of the alphabetical index included 49,097 lines. After the processing, the new file with codes, tags and terms now includes 44,559 codes with 10,252 supplementary words. In addition, two separate files with 3,822 cross-references "see" and 2,204 cross-references "see also" have been created.

The initial application included 14,475 main terms corresponding to the 14,475 ICD10 codes. By breaking into component parts the inclusions, 36,873 have been added and the integration of the alphabetical index has brought 42,046 supplementary strings.

The initial thesaurus was composed of 22,304 words and 3,358 words were added. The links between codes and strings have been obviously increased (nearly 100,000 added links), which has led to a decrease of documentary silence.

Discussion

The suggested theoretical model allows to take into account inclusions, exclusions and hierarchical relations between codes. However, it is probably inadequate to represent more sophisticated notions included in the ICD10 like coding an underlying disease or a manifestation (dagger and asterisk codes), an action ("radiotherapy session") or a social problem ("problems related to unemployment"). For these complex notions a multi-axial knowledge representation would be more relevant.

The theoretical modeling of the ICD under the object form would enable an object-oriented database. But up to now, this has not been possible on micro-computers. The aim was to offer an easy-usable tool to most medical practitioners that is why the implementation on micro-computer has been kept. Therefore, the theoretical model described has been considerably simplified.

The increase of the number of key words and of links between words and codes have lessened documentary silence. But by doing so, documentary noise has been increased, which led to interference in some researches and to non relevant codes. Up to now, the addition of the alphabetical index does not clearly demonstrate whether it entails an increase or a decrease of the coding help system performances since it has not been possible to clearly evaluate them.

Compared to the numerous supplementary strings, the small number of new words may be surprising. However, the thesaurus was initially composed of words coming from the analytical ICD10 as well as words from other classifications not linked to the ICD codes. Besides, many of these added strings are likely to be lexical variants of pre-existing strings.

A qualitative analysis of the added word list has shown that the new words were either simple words or eponyms (Eberth, Faber) or compound-words, a part of which were already present in the thesaurus. The use of more elaborated linguistic tools would certainly improve the relevance of indexation where compound-words or lexical variants are concerned.

This attempt to an automatic processing of the ICD strings has enlightened the fact that informatic tools are difficult to operate to transform a classification primarily designed to be manually used. For example it was impossible to separate automatically synonyms and included strings of a main term. Some precisions put into parentheses have also been difficult to adapt. The initial project was to replace strings with parenthesis-included precisions by different strings without parentheses. However, the analysis of these precisions have shown that they were logically associated either with an "exclusive or" or with an "inclusive or", and such a differentiation could not be automatic. e.g.:

Fracture ankle (bimalleolar) (trimalleolar)

These supplementary words are associated with an exclusive or (XOR) and should lead to 2 different strings.

Poliomyelitis (acute) (anterior) (epidemic)

This example shows an association with an inclusive or (OR).

Conclusion

The attempt to integrate the alphabetical index into the coding help system in use in our department has led to a study of the ICD structure. The alphabetical index is composed of many strings that do not appear in the analytical index; its integration has increased the research possibilities. However, since this classification is meant to be manually used, the automatic processing has aroused an amount of difficulties. Modifying each string separately could only solve some of them. Therefore, these modifications will have to be carried out whenever the classification is updated, which is repeatedly done with the French version used in the hospitals. The setting of a tool with independence between program and data seems to be difficult.

A lot of work has still to be performed to obtain an effective medical coding help system using the ICD10. Nevertheless, such a system will remain an intermediary step requiring human acting to select diagnoses. The ultimate step would be a completely automatic coding system using the natural language processing.

References

[1] International Statistical Classification of Diseases and Related Health Problems. World Health Organisation, Geneva, 1992; volumes 1 and 3.

[2] Landais P, Jais JP, Frutiger P. Sémantique des classifications et nomenclatures. In Informatique et Santé. Springer-Verlag 1989;pp 211-22.

[3] UMLS knowledge sources. 5th Experimental Edition April 1994 - documentation. National Library of Medicine.

[4] McCray AT, Nelson SJ. The representation of meaning in the UMLS. Meth Inform Med 1995; 34:193-201.

[5] Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of a large controlled medical terminology. J Am Med Informatics Assoc 1994; 1(1): 35-50.

[6] Genillier PL, Prevel M, Lescanne FL. Mise en place d'un système d'aide au codage. Informatique et santé - La revue 1993; 15:11-4.

[7] Moore GW, Berman JJ. Performance Analysis of manual and automated systemized nomenclature of medicine (SNOMED) coding. Am J Clin Pathol 1994; 101(3): 253-6.

[8] Baud R, Lovis C, Rassinoux AM, Scherrer JR. Tendances en traitement du langage naturel. In: Informatique et Santé. Paris: Springer-Verlag 1996;pp 111-9.

[9] Baud RH, Lovis C, Alpay L, Rassinoux AM, Scherrer JR, Nowlan A, Rector A. Modelling for natural language understanding. In: Safran C (ed). Proceedings SCAMC 1993. New York: McGrawHill, 1993;pp 289-93.

[10] Cimino JJ. Coding systems in health care. Yearbook of medical informatics 1995:71-85.

[11] Kohler F, Mayeux D, Musse JP. Informatique et aide au codage du PMSI. Gestions Hospitalieres 1990;299:662-6.

Address for correspondence

Christine Bouchet. Laboratoire
SPI-EAO - Faculté de Médecine de Nancy
BP 184 - 54505 Vandoeuvre Cedex
France.
bouchet@spieao.u-nancy.fr