MEDINFO 2010
C. Safran et al. (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-588-4-1080

Development of Structured ICD-10 and its Application to Computer-Assisted ICD Coding

Takeshi Imai[a], Masayuki Kajino[b], Megumi Sato[b], Kazuhiko Ohe[c]

[a] *Center for Disease Biology and Integrative Medicine, Graduate School of Medicine, The University of Tokyo, Japan*
[b] *Department of Planning, Information and Management, The University of Tokyo Hospital, Japan*
[c] *Department of Medical Informatics, Graduate School of Medicine, The University of Tokyo, Japan*

Abstract

This paper presents: (1) a framework of formal representation of ICD10, which functions as a bridge between ontological information and natural language expressions; and (2) a methodology to use formally described ICD10 for computer-assisted ICD coding. First, we analyzed and structurized the meanings of categories in 15 chapters of ICD10. Then we expanded the structured ICD10 (S-ICD10) by adding subordinate concepts and labels derived from Japanese Standard Disease Names. The information model to describe formal representation was refined repeatedly. The resultant model includes 74 types of semantic links. We also developed an ICD coding module based on S-ICD10 and a 'Coding Principle,' which achieved high accuracy (>70%) for four chapters. These results not only demonstrate the basic feasibility of our coding framework but might also inform the development of the information model for formal description framework in the ICD11 revision.

Keywords

ICD10, Ontology, Knowledge bases, Natural language processing, Computer-assisted coding

Introduction

The World Health Organization (WHO) officially launched the 11th revision of the International Classification of Disease (ICD) in April 2007 [1]. One important planned feature of ICD11 is structurizaton of the clinical meaning of each disease category to provide formal representations of ICD categories to describe the characteristics of each disease in various dimensions such as Etiology, Anatomic site, Manifestation attributes, and Pathophysiology. Figure 1 shows a possible formal representation of 'Venezuelan equine encephalitis (A92.2).' The meaning of the disease concept is represented as a tree-structure using two component concepts (*CC*) (e.g "Encephalitis") and semantic links of two types (e.g. "<hasCause>").

The structurization process might have several levels of granularity, but such a formal representation of ICD will be useful for advanced information retrieval systems. It is anticipated as a useful knowledge base for computer-assisted ICD coding.

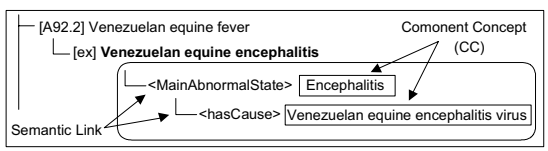

Figure 1 – Simple example of 'formal representation'.

Computer-assisted ICD coding (or Automated ICD coding) has attracted attention continuously since the 1990s. Although several studies have sought to represent ICD categories formally and to use it for ICD coding [2–8], their framework and information models for formal representations were insufficient to cover all ICD categories because they investigated only a few domains. For example, Héja et al. [6,7] developed an information model to describe ICD categories using six concept categories and semantic links of five types. However, they analyzed only two chapters. Moreover, it is not clear that such a small information model is applicable to all chapters. No comprehensive information model has included the complete list of semantic links to describe all ICD categories.

In addition, the ICD coding frameworks in those studies did not consider distinctions among concepts and their string expressions (labels). Fundamentally, the first step to perform ICD coding is to map input strings to *CCs*. However, that is not always possible because of an omission or an abbreviation. As might be apparent, no substring of the input disease name "*Venezuelan equine encephalitis*" can be mapped directly to the CC in Fig. 1 – "*Venezuelan equine encephalitis virus.*" A new methodology to address mapping between concepts and string expressions is needed to improve coding results.

This study has two major goals. The first is to develop a structured ICD10 ('*S-ICD10*'), which functions as a bridge between ontological information and natural language expressions, based on a robust and comprehensive information model that can cover all ICD categories. The second is to develop a methodology to use S-ICD10 for computer-assisted ICD coding.

Materials and Methods

Development of S-ICD10

Step 1): Structurizing the ICD10 Tabular List

First, 20 Japanese ICD coders manually analyzed the ICD-10 book (Volume 1: Tabular List, 2003 Japanese edition) and described formal representations for all ICD categories and example entries in 15 chapters (excluding Chps. 5, 15, 16, 18, 20, 21, and 22). It is noteworthy that Chps. 20, 21 and 22 are additional information, so the number of main chapters in ICD10 is 19. As Fig. 2 shows, each ICD category and an example entry has at least one formal concept representation (hereinafter '*FCR'*) represented as a tree structure, and each semantic link has cardinality information. The main tasks in this step were: (1) to identify CCs from the title of each category; and (2) to assign semantic links to CCs to form tree representations. Two Japanese medical and ontology experts reviewed all results. If at least one disagreed with the result, then the information model and the list of all semantic links were reconsidered; all descriptions were revised based on the new information model. We started this project in 2005. These iterative revisions were repeated three times.

Step 2): Expanding S-ICD10 and adding labels

Step 1 includes no distinction between CCs and their labels. Therefore, in this step, we separated those labels from CCs and assigned additional labels derived from Japanese Standard Disease Names (*JSDN*) [9] to S-ICD10. The 25,280 disease names in JSDN were manually parsed (every disease entry in JSDN has a proper ICD10 code); the 20 annotators (same as step 1) assigned each token—a morpheme or substring of each disease name—to the corresponding FCR as one of the following types: (1) *a direct label of CC(s)*; (2) *a label of subordinate concept of CC(s)*; and (3) *a label of an additional CC of the FCR.* For example, in Fig. 1, we assigned the label '*Venezuelan equine*', as a direct label of CC *[Venezuelan equine encephalitis Virus]*. It means that the string '*Venezuelan equine*' can indicate the CC under a certain context, even though the string itself is not a virus name. Regarding M254 in Fig. 4, a new concept *'[Knee Joint]'* together with its label '*Knee Joint*' was added to the CC *[Joint]*, as its subordinate concept; a new concept '*[Swelling]*' together with its label was also added to the FCR as an additional CC with cardinality '0..1'. These new concepts were derived from disease entries in JSDN that have the 'M254' ICD code.

This step was very important to perform ICD coding. ICD10 is a classification system and each ICD category is an aggregation of diseases. Therefore, component concepts of an input disease are sometimes more granular than CCs of ICD entries. Two experts (same as step1) reviewed the results obtained in this step. Furthermore, in cases where they did not agree, the annotation result was excluded. Results show that approximately 85% of all tokens in JSDN were included in the expanded layer of S-ICD10.

Automated ICD coding framework based on S-ICD10

Overview of our coding framework

Figure 2 depicts an overview of our coding framework. The coding module leverages a Japanese general tagger called YOMOGI, which we developed in 2007, for tokenizing an input disease name based on the label set in S-ICD10.

First, YOMOGI outputs *N*-best tokenization of an input disease name. Each token has corresponding CC(s) in S-ICD10. However, some tokens might correspond to various CCs in different ICD categories. For example, as shown in Fig. 2, an input disease name was tokenized into four tokens, and the token '*universal*' corresponds to a label of the CC in B007, D65, K650, L631, and so on. The system then considers all possible combinations of corresponding CCs and selects one which best covers a certain ICD code. As Fig. 2 shows, the system regarded "*universal*" as a label of CC in D65:ex3/ D65 and "*blood*" as a label of CC in D50–D89, because both D65 and D50–D89 are upper categories of D65:ex3.

Finally, the input disease is mapped to the ICD entry 'D65:ex3'; the system then outputs its ICD code (D65).

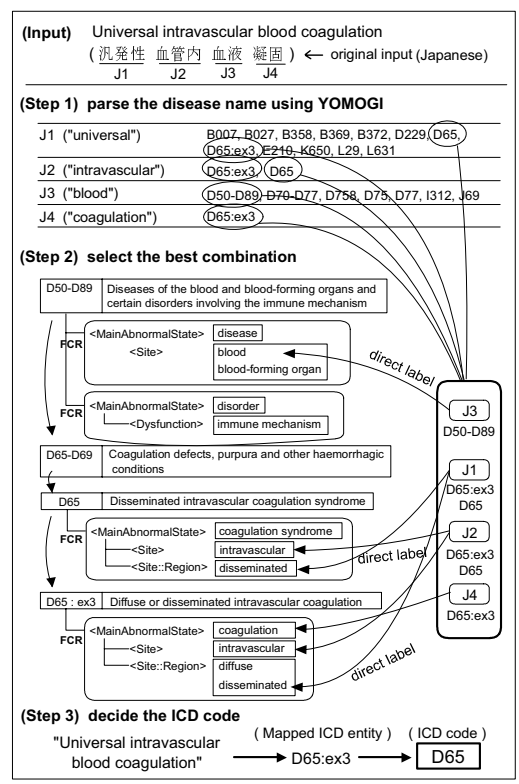

Figure 2 – Overview of our coding framework.

'Coding principle'

In the coding process, the system uses a '*Coding principle*' to decide ICD codes. From the ontological perspective, 'Concept B' is a child concept of 'Concept A' if: (1) every CC in 'Concept B' is the same as or the specialization of the corresponding CC in 'Concept A'; or (2) some additional CCs exist aside from the condition explained above, as shown in Figure 3.

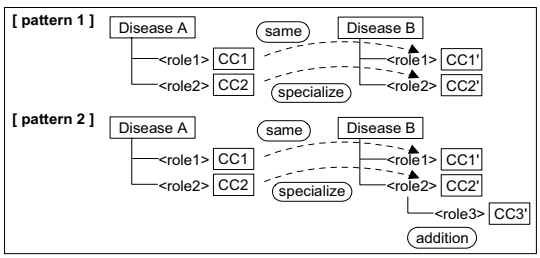

Figure 3 – Is-a relations between two concepts.

However, those rules are too strict for ICD coding purposes. In Fig. 4, the "*swelling*" label in the input string was mapped to the CC in M254; the "*knee joint*" label was mapped to the subordinate concept of the CC in M254. However, not all CCs in M254 were covered by the input string, although the input disease concept is fundamentally a child concept of M254. Therefore, we cannot apply the [pattern1] rule in Fig. 3 to this case. We used a *'Coding principle'* to solve this problem—"An input disease has the ICD code 'X' if every token in the input disease can be mapped to: (1) CC in X; (2) subordinate concept of CC in X; or (3) CC that can be inherited from ancestor categories of X". This '*Coding principle'* does not require that all CCs in X whose cardinality is one or more be covered by the input disease name. In that sense, it is a weakened condition of two patterns in Figure. 3 for coding purposes.

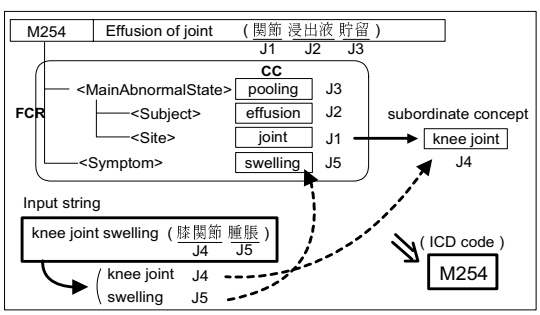

Figure 4 – Label mapped to a subordinate concept of CC. (where J1-J5 denote Japanese tokens)

If an input disease name cannot satisfy the Coding Principle, then the system outputs ICD code candidates according to the coverage ratio (# of tokens satisfying the Coding Principle / # of total tokens). The system outputs nothing in cases where the coverage ratio is low (< 50%).

Evaluation on the ability for ICD coding

To investigate the performance of the proposed coding framework, we evaluated the capability for ICD coding based on S-ICD10 and the '*Coding principle*'. We randomly chose 1,255 disease names collected from various hospitals in Japan and coded them manually. Entries present in Japanese Standard Disease Names (JSDN) were excluded beforehand because S-ICD10 had already included knowledge derived from JSDN.

Results

S-ICD10 and the information model to describe FCRs

Table 1 – Categorization of all semantic links

Category	Type of semantic link	# of ST	Freq.
Pathophysiology	Main Abnormal State	-	13,111
Site	Has Site	2	8,411
	Related Site	-	54
Cause	Cause	1	4,146
	Modifier of Cause Entity	2	4
	Cause-related	5	91
Function	Dysfunction	-	247
Temporal Relation	Features of Occurrence	-	298
	Progress/ Timing/ Age	-	1,110
Symptoms/ Findings	Symptoms/ Findings	1	1,428
	Modifier of Symptoms/ Findings	6	71
Specification of other items	State	11	545
	Treatment-related	2	2
	Type	3	152
	Route/ Mode of intake	2	10
	Certainty	3	3
	Others	7	35
Subject of others	Subject	2	565
Examination/ Diagnosis	Mode of confirmation	-	81
	Diagnostic method	-	1
Relation to other disorders	-	11	1,899
Other	-	9	66

15,221 ICD entries (categories and examples) in 15 chapters were structured. S-ICD10 has 15,463 FCRs, 55,453 CCs (20,320 unique CCs), and 81,478 labels (39,164 unique labels) in total. The information model to describe FCR includes 74 types of semantic links. Table 1 shows a categorization of all semantic links.

Each type of semantic link might have sub-types (the third column shows the number of sub-types). For example, the type 'State' includes 11 sub-types of semantic links, such as 'Shape', 'Benign/Malignant', 'Atypia', 'Severity', 'Quantity', and 'Grade of Progress'. The category 'Relations to other disorders' includes 11 sub-types, such as 'Complication', 'Underlying disease', 'Follow', and 'Sequela'. The fourth column shows the frequency of the semantic link. For example, the semantic link 'Diagnostic method' was used only once in the formal representation of 'I252:ex2' – "*Past myocardial infarction diagnosed by ECG or other special investigation, but currently presenting no symptoms.*"

Automated ICD Coding

Table 2 shows results of automated ICD coding. Overall, 61.7% (747/1211) disease names were coded correctly. The best result (76.4%) was Chapter 7 (Diseases of the eye and adnexa); the worst (46.2%) was Chapter 4 (Endocrine, nutritional and metabolic diseases). We excluded diseases of four chapters (5, 15, 16, and 18) because S-ICD10 does not cover them. However, diseases from those excluded chapters were only 44 (total = 1,255). Therefore, if those categories are included later, they will little affect the overall result.

Table 2 – Automated coding results

Ch	#C	#N	*R*(%)	Ch	#C	#N	*R*(%)
01	72	28	72.0	10	32	13	71.1
02	61	56	52.1	11	35	30	53.9
03	19	20	48.7	12	109	47	69.9
04	42	49	**46.2**	13	37	30	55.2
06	32	12	72.7	14	46	24	65.7
07	29	9	**76.3**	17	26	17	60.5
08	15	8	65.2	19	144	82	63.7
09	48	39	55.2	**Total**			**61.7**

Note: Ch, Chapter No.; **#C**, # of correctly coded diseases; **#N**, # of non-coded diseases; ***R*(%)**, the ratio of #C).

Discussion

Disease description framework

The S-ICD10 result description process showed that all ICD entries, at least in 15 chapters (of 19 main ones), can be represented formally using our information model to describe FCR. The information model based on the categorized list of the semantic link types shown in Table 1 is much more granular than in any previous study. Many newly found semantic link types were indispensable for formal representation of ICD categories, showing the importance of comprehensive analysis.

Some semantic links must be refined for further development toward more sophisticated ontological representation of ICD10. Also, a '*Subordinate'* relation should be separated into '*Is-a*' and '*Part-of*' relations from an ontological perspective. Consequently, we call it '*Structured*' ICD-10, not '*Ontology*'.

Nevertheless, the information model is useful for development toward ontological representation. It might inform the ICD11 revision project as a pilot study to create a comprehensive information model to describe ICD categories formally.

Computer-assisted ICD coding

As for the ability for ICD coding, although it is difficult to compare results among coding systems which use different test sets, the overall accuracy (61.7%) is similar to the best result among previous studies [13] and much better than other previous results. The system achieved high accuracy (>70%) for four chapters, but the accuracies of Chapter 3 and 4 diseases were low. The main reason is the lack of subordinate concepts, especially in anatomical entities. We added many subordinate concepts and labels derived from JSDN. However, JSDN had insufficient information to cover all disease names collected from various hospitals in Japan. A possible solution is the use of semantic relations between anatomical entities defined in other ontologies such as FMA and SNOMED-CT.

The system outputs other ICD codes along with the correct one in cases where candidates have equal scores (coverage ratios). We counted those cases as 'correctly coded' in the evaluation study because it is apparently easy for human coders to select the correct code from those candidates in later screening. The system can also output the reason for the coded result, showing the mapping result between tokens in the input disease names and CCs, which will be helpful for later screening.

The overall accuracy decreases to 34.8% if we do not use the 'Coding Principle' and perform ICD coding based on simple matching between an input and FCR. The 'Coding Principle' is important in terms of: (1) using information (CCs) inherited from the upper categories: and (2) allowing the case in which not all CCs of a certain FCR are covered. This important feature of our coding framework enables us to address the case in Fig. 4, which the previous knowledge representation-based approaches could not cover.

Related works

Substantial efforts have been made for automated or computer-assisted ICD coding to date. They are classifiable into two types: (A) using disease names or clinical notes, which already have correct ICD codes, to calculate similarities by statistic measure (***'example-based'***) [10–13]; and (B) using formal representation of ICD categories to describe coding rules (***'knowledge representation-based'***) [2–8].

Type-A methods can be implemented easily, but they have not shown high accuracy. The coding systems require numerous examples to achieve better results. However, it is difficult to collect them evenly. Some ICD codes have no coded example.

Moreover, it cannot provide explanation capability, which is useful for later screening by human coders.

On the other hand, Type-B methods present advantages to provide explanation capabilities. However, developers must describe vast amounts of knowledge. Therefore, previous studies only proposed designed framework or implemented the system in a small limited domain. Our framework is also 'knowledge representation-based'. However, it differs from other Type-B studies in the following respects: (1) our project achieved high coverage (15 of 19 main chapters), and our information model representing ICD10 categories is considered highly robust; (2) 'concepts' and their 'labels' are distinguished explicitly so that S-ICD10 can work as a bridge between ontological information and natural language expressions; (3) our coding framework uses the 'coding principle', which allows property inheritance and weakened conditions of concept subsumption.

Limitations and future directions

The S-ICD10 covers most chapters, but some, such as 'Mental and behavioral disorders (Chap. 5)' have not been addressed. The information model to represent disease concepts might not be sufficient to cover all disease concepts in ICD10. We plan to apply our methodology to the remaining four chapters and to produce a more comprehensive formal representation of ICD10. We also plan to map all concepts (CCs) and labels to the current existing terminologies and ontologies such as SNOMED-CT, GALAN, FMA, and Japanese Medical Ontology, and to create an English version of S-ICD10.

Conclusion

This paper presents: (1) a framework of formal representation of ICD-10, which functions as a bridge between ontological information and natural language expressions; and (2) a methodology to use S-ICD10 and the '*Coding Principle*' for computer-assisted coding. The results demonstrate the effectiveness of our framework. In fact, S-ICD10 has unprecedentedly high coverage of all ICD10 categories. The resultant information model to describe formal representation of ICD categories might inform ICD11 revision as a pilot study.

Acknowledgments

This research was supported by the Ministry of Health, Labour and Welfare of the Japanese Government as Development and Research of "Medical-knowledge-based database for medical informatics system," and by a Grant-in-Aid for Young Scientists (B) (19700128) from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

[1] Revision of the International Classification of Diseases: http://www.who.int/classifications/icd/ICDRevision/en/

[2] Delamarre D, Burgun A, Seka LP, Le Beux P. Automated coding of patient discharge summaries using conceptual graphs. Methods Inf Med. 1995;4(34):345-51.

[3] Bernauer J, Schoop D. Formal classification of medical concept descriptions: graph-oriented operators. Methods Inf Med. 1998 Nov;37(4-5):510-7.

[4] Bouchet C, Bodenreider O, Kohler F. Integration of the analytical and alphabetical ICD10 in a coding help system. Proposal of a theoretical model for the ICD representation. Stud Health Technol Inform. 1998;52 Pt 1:176-9.

[5] Fabry P, Baud R, Ruch P, Le Beux P, Lovis C. A frame-based representation of ICD-10. Stud Health Technol Inform. 2003;95:433-8.

[6] Héja G, Surja'n G, Luka'csy G, Pallinger P, Gergely M. GALEN based formal representation of ICD10. Int J Med Inform. 2007 Feb–Mar;76(2-3):118-23.

[7] Héja G, Varga P, Surján G. Design principles of DOLCE-based formal representation of ICD10. Stud Health Technol Inform. 2008;136:821-6.

[8] Jiang G, Pathak J, Chute CG. Formalizing ICD coding rules using Formal Concept Analysis. J Biomed Inform. 2009 Jun;42(3):504-17.

[9] Japanese Standard Disease Names: http://www.dis.h.u-tokyo.ac.jp/byomei/

[10] Michel PA, Lovis C, Baud R. LUCID: a semi-automated ICD-9 encoding system. Medinfo. 1995;8 Pt 2:1656.

[11] Pakhomov SV, Buntrock JD, Chute CG. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. J Am Med Inform Assoc. 2006 Sep–Oct;13(5):516-25.

[12] Tagliabue G, Maghini A, Fabiano S, Tittarelli A, Frassoldi E, Costa E, Nobile S, Codazzi T, Crosignani P, Tessandori R, Contiero P. Consistency and accuracy of diagnostic cancer codes generated by automated registration: comparison with manual registration. Popul Health Metr 2006;4:10.

[13] Aramaki E, Imai T, Kajino M, Miyo K, Ohe K. Statistical selector of the best multiple ICD-coding method. Stud Health Technol Inform. 2007;129(Pt 1):645-9.

Address for correspondence

Takeshi IMAI: Center for Disease Biology and Integrative Medicine, Graduate School of Medicine, The University of Tokyo, Japan.
7-3-1 Hongo, Bunkyo, Tokyo 113-8655, Japan.
E-mail: ken@hcc.h.u-tokyo.ac.jp