Development of ICD-10-TM Ontology for a Semi-automated Morbidity Coding System in Thailand

S. Nitsuwat; W. Paoin
Faculty of Information Technology, King Mongkut’s University of Technology North Bangkok (KMUTNB), Thailand

Keywords
ICD-10 ontology, semi-automated ICD coding system, ICD knowledge base, ICD search

Summary
Objectives: The International Classification of Diseases and Related Health Problems, 10th Revision, Thai Modification (ICD-10-TM) ontology is a knowledge base created from the Thai modification of the World Health Organization International Classification of Diseases and Related Health Problems, 10th Revision. The objectives of this research were to develop the ICD-10-TM ontology as a knowledge base for use in a semi-automated ICD coding system and to test the usability of this system.

Methods: ICD concepts and relations were identified from a tabular list and alphabetical indexes. An ICD-10-TM ontology was defined in the resource description framework (RDF), notation-3 (N3) format. All ICD-10-TM contents available as Microsoft Word documents were transformed into N3 format using Python scripts. Final RDF files were validated by ICD experts. The ontology was implemented as a knowledge base by using a novel semi-automated ICD coding system. Evaluation of usability was performed by a survey of forty volunteer users.

Results: The ICD-10-TM ontology consists of two main knowledge bases (a tabular list knowledge base and an index knowledge base) containing a total of 309,985 concepts and 162,092 relations. The tabular list knowledge base can be divided into an upper level ontology, which defines hierarchical relationships between 22 ICD chapters, and a lower level ontology which defines relations between chapters, blocks, categories, rubrics and basic elements (include, exclude, synonym etc.) of the ICD tabular list. The index knowledge base describes relations between keywords, modifiers in general format and a table format of the ICD index. In this research, the creation of an ICD index ontology revealed interesting findings on problems with the current ICD index structure. One problem with the current structure is that it defines conditions that complicate pregnancy and perinatal conditions on the same hierarchical level as organ system diseases. This could mislead a coding algorithm into a wrong selection of ICD code. To prevent these coding errors by an algorithm, the ICD-10-TM index structure was modified by raising conditions complicating pregnancy and perinatal conditions into a higher hierarchical level of the index knowledge base. The modified ICD-10-TM ontology was implemented as a knowledge base in semi-automated ICD-10-TM coding software. A survey of users of the software revealed a high percentage of correct results obtained from ontology searches (>95%) and user satisfaction on the usability of the ontology.

Conclusion: The ICD-10-TM ontology is the first ICD-10 ontology with a comprehensive description of all concepts and relations in an ICD-10-TM tabular list and alphabetical index. A researcher developing an automated ICD coding system should be aware of the ICD index structure and the complexity of coding processes. These coding systems are not a word matching process. ICD-10 ontology should be used as a knowledge base in the ICD coding software. It can be used to facilitate successful implementation of ICD in developing countries, especially in those countries which do not have an adequate number of competent ICD coders.

1. Introduction
The new era of health information systems and interactive web technology in modern health care facilities has introduced a number of research and development issues whereby the efficient integration of disease classification knowledge with automated coding process is needed. In the last decade the application of ontologies in information systems was promoted [1] as a shared platform for information integration.
The International Classification of Diseases and Related Health Problems, 10th Revision – ICD-10 is a classification that was created by the World Health Organization (WHO) in 1992 [2] and has been maintained by them since then. Electronic versions of ICD-10 were released in 2004 as browsing software in a CD-ROM package [3] and as an online ICD-10 tool on the WHO website [4]. Both of these electronic versions provided only a simple word search service that facilitated only a minor part of the complex ICD coding process. Since 2000, some countries, e.g., Australia,
Canada, Germany etc., have used medical expert opinions to add more codes into ICD-10. In Thailand, the tabular list (volume 1) of ICD-10 was translated into the Thai language since 2000. In 2003 ICD-10 was modified to be ICD-10-TM (Thai Modification) [5]. ICD-10-TM had 1,072 more codes than the original ICD-10, most of new codes were created from list of diseases found in Thailand but were not included in the original ICD-10, for example – green pit viper snake bite, dengue hemorrhagic shock syndrome, pruritic popular eruption etc. The added ICD-10-TM codes were used in 5–10% of Thailand mortality coding cases [6].

In this paper, we present an ICD-10-TM semi-automated coding platform that assists the ICD coding process by combining the use of the ICD-10-TM ontology and a novel ontology search methodology. The semi-automated coding model developed in this work can be used as a basic infrastructure to build a web service or a coding software package that can help ICD coders to assign ICD code in a more effective way.

### 1.1 Ontology Evolution

A study [7] indicates that in the next ten years an integrated ‘Semantic Web’ will be one of the emerging technologies that will have the highest impact on the way we interact with the internet. Ontologies are used as a foundation to enable interoperability within the Semantic Web. Ontology is a data structure and data representation tool that can be used to share and reuse knowledge between artificial intelligence systems which share a common vocabulary. In the medical domain, researchers and health care standard bodies have started to use ontologies to share their data and knowledge.

Research in ontology evolution encompasses foundational work that focuses on ontology creation and manipulation techniques. The ultimate goal of these techniques is ontology reuse and integration. Hu et al. [8] have classified ontology creation into top-down and bottom-up methods. Arenas et al. [9] have suggested the enhancement of the bottom-up building of ontology by creating several small ontologies and then merging them together into a large ontology by using automatic merging software. Rector [10] proposed that modularization of ontology implementation could facilitate merging, alignment, maintenance and evaluation processes for any ontology. Various studies (see, e.g., [11, 12]) have also used ontologies for creating intelligent systems.

### 1.2 Application of Ontologies in the Medical Domain

Yu [13] discussed three ways of applying ontologies in the medical domain, namely, i) terminology management, ii) integration, interoperability and sharing of data, and iii) knowledge reuse and decision support. Cimino’s work [14] on a Medical Entity Dictionary used in the New York Presbyterian hospital was one of the pioneer works in the application of ontology techniques in the management of medical terminology. While the System Nomenclature of Medicine - Clinical Terminology (SNOMED-CT) [15] was created long before the development of modern ontology science, several studies [16, 17] suggested modification of the terminology so that it complied with current standards of ontology. Likewise, the Unified Medical Language System – UMLS [18] developed by the United States Library of Medicine before the era of modern ontology, was modified recently to be presented as a semantic network and to create a new upper-level ontology for use in the medical domain [19]. Although there were some problems, recent work by Pisaneli [20] showed an effort to adapt current ontology to be a framework for representing medical terminological knowledge.

Kumar et al. [21] used an ontology created in the UMLS framework to share data use in a colon cancer research project. Orgun et al. [22] used the ontology created from HL7 standard and mobile agents.

### Table 1: Some ICD concepts and relations from the tabular list and alphabetical indexes in RDF N3 format

<table>
<thead>
<tr>
<th>ICD Entity</th>
<th>Ontology type</th>
<th>RDF format</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter</td>
<td>concept</td>
<td>icd10:Chapter</td>
<td>icd10:Chapter1</td>
</tr>
<tr>
<td>Block</td>
<td>concept</td>
<td>icd10:Block</td>
<td>icd10:BlockA00A09</td>
</tr>
<tr>
<td>Category</td>
<td>concept</td>
<td>icd10:Cat</td>
<td>icd10:CatA00</td>
</tr>
<tr>
<td>Code</td>
<td>concept</td>
<td>icd10:</td>
<td>icd10:A000</td>
</tr>
<tr>
<td>Term</td>
<td>concept</td>
<td>term:</td>
<td>term:Hemorrhoids</td>
</tr>
<tr>
<td>hasCode</td>
<td>relation</td>
<td>icd10:hasCode</td>
<td>icd10:CatA00 icd10:hasCode icd10:A000</td>
</tr>
<tr>
<td>hasLabel</td>
<td>relation</td>
<td>icd10:hasLabel</td>
<td>icd10:CodeA00 icd10:hasLabel term:Cholera_due_to_Vibrio_01</td>
</tr>
<tr>
<td>hasChildTerm</td>
<td>relation</td>
<td>icd10:hasChildTerm</td>
<td>term:Hemorrhoids icd10:hasChildTerm term:External_Hemorrhoids</td>
</tr>
<tr>
<td>hasInclude</td>
<td>relation</td>
<td>icd10:hasInclude</td>
<td>icd10:Chapter1 icd10:hasInclude term:Disease_generally_recog…</td>
</tr>
<tr>
<td>hasExclude</td>
<td>relation</td>
<td>icd10:hasExclude</td>
<td>icd10:Chapter1 icd10:hasExclude term:Carrier_or_suspected..</td>
</tr>
<tr>
<td>pointto</td>
<td>relation</td>
<td>index:pointto</td>
<td>term:Hemorrhoids index:pointto icd10:1849</td>
</tr>
</tbody>
</table>
The ICD-10-TM upper level ontology. Chapter 21 was placed in the first node after the root node (all conditions). Chapters 15 and 16 were classified as chapter for specific population. Chapter 1, 2, 17, 19 and 20 were etiology specific chapters and were in a higher level in the ontology than diseases classified by organ chapter. Chapter 18 was in the lowest node since sign and symptom codes would only be used in cases where no diagnosis could be made.

Fig. 1

1.3 The ICD Coding Processes

The ICD coding processes have been summarized in ICD-10 volume-2 [27]. This volume proposes seven steps that should be followed before assigning an ICD code. These steps are: i) identify the diagnosis statement to be coded; ii) locate the lead term in the alphabetical index; iii) read and be guided by any note that appears under the lead term; iv) read any terms indented under the lead term until all the words in the diagnostic expression have been accounted for; v) follow carefully any cross-references found in the index; vi) refer to the tabular list to verify the suitability of the code selected; and vii) be guided by any inclusion or exclusion terms under the selected code or under the chapter, block or category heading. Clinical coders always need at least two ICD books to complete all the coding processes. These books are the tabular list and the index. The ICD coding processes are complex and difficult to carry out manually. The ICD coding can not be finished merely by word matching between diagnosis words and a list of ICD codes/labels. Unfortunately, this complexity of ICD-coding was not recognized by most researchers who have tried to develop semi-automated and automated ICD coding systems in the past. These systems used diagnosis word matching process only, so the systems did not include the last two WHO ICD coding steps ie. (vi – verify code after word matching) and (vii – check inclusion and exclusion terms) which are very important for accurate coding. For example, a...
Fig. 2  Example of the lower level tabular list ontology structure of the ICD category I84 haemorrhoids
patient with a car accident was diagnosed with word “intracerebral hemorrhage” will be code by the system as I61.9 (intracerebral hemorrhage) while a coder who does manual coding will select the code S06.3 (traumatic intracerebral hemorrhage) after checking exclusion terms under the chapter 9 disease of the circulatory system (100-199). The Yamada study [28] showed a high ICD coding error rate (32%) generated from a coding system used by physicians. These errors could be avoided by manual coding done by coders. Incorrect ICD coding could lead to quality problems in morbidity and mortality statistics which

---

Fig. 3  Example of ICD-10-TM index ontology, keyword = hemorrhoids. This ontology structure was created after expert modification of ICD-10-TM index.
affect health prevention program prioritization in every country.

1.4 Semi-automated and Automated ICD Coding Systems

Representation of ICD-10 in electronic format started in 1998 by the Australian National Centre for Classification in Health (NCCH) [29]. The NCCH converted the classification of ICD-10-AM (Australian Modification) to a relational database. The database was used to develop further electronic coding products as well as to make the classification available in electronic format to users and software developers. ICD-10-AM’s second edition was printed from this database in 2000.

Several authors have described automated ICD coding processes in their papers. Lovis et al. [30] described a Diogene 2 program designed to build a medical terminology table. They used it to map a diagnosis word into a form/meaning layer, then to convert the corresponding term into a concept layer before matching it to labels of ICD codes in an expression layer. Heja et al. [31] compared several models based on N-grams, vector space models, and neural networks for matching diagnosis words with a list of ICD code labels and suggested that a hybrid model yielded better matching results. Pakhomov et al. [32] designed an automated coding system that assigned codes to out-patient diagnoses using example-based and machine learning techniques. Periera et al. [33] built a semi-automated coding help system using an automated MeSH-based indexing system and a mapping between MeSH and ICD-10 extracted from the UMLS metathesaurus. However, these previous work only used processes of word matching and did not cover full standard ICD coding processes.

The ICD is not a disease terminology system. Research work in the past did not aware about this fact. Our work will contribute a novel system that uses the ICD ontology as a knowledge base for development of semi-automated ICD coding systems. This ICD ontology is a data structure that represents knowledge on classification of disease. The ontology could then be reused in new sys-
tems that require knowledge about the ICD without the need for undertaking additional development work.

2. Objectives

The objectives of this study are 1) to create the ICD-10-TM ontology as a knowledge base for a semi-automated ICD coding system; and 2) to test usability of the ICD-10-TM ontology in a semi-automated ICD coding system.

3. Methods

All ICD concepts and relations from the tabular list and alphabetical indexes were identified (Table 1). The ontology was defined in resource description framework (RDF) model (N-triples format-N3) [34]. ICD-10-TM contents in Microsoft Word documents were transformed into N3 format using scripts written in the Python 3.1 script language. The final RDF files were validated by human experts using Cytoscape for graph data visualization and nodes count. The ontology was implemented as a knowledge base used by a novel semi-automated ICD coding system written in the JAVA Netbeans 6.9 platform. Forty volunteer users then evaluated the usability of the ontology in the system.

4. Results

The ICD-10-TM ontology contains 309,985 concepts and 162,092 relations. The ontologies stored its knowledge in two major sections, the tabular list ontology and the index ontology.

4.1 Tabular List Ontology

The tabular list ontology represents classification schemas and can be divided into an upper level ontology, which represents relationships between ICD chapters (Fig. 1), and a lower level ontology which represents relationships within each ICD chapter i.e. relationships between chapter and blocks, block and other block, categories and rubric (Fig. 2). Relationships specified in the upper level ontology are parent and child relationships. In Figure 1, the root node (all conditions) has a chapter 21 (factor influencing health status and health service) as a child node. The node “illnesses needed for treatment” has three children, i.e., diseases found in pregnancy, diseases found in the neonatal period and diseases found in general population.

There are many types of relationships specified in the lower level ontology of the tabular list, including “parent of”, “name as”, includes, excludes, “label as”, synonym etc. In an ontology graph, relationships would be shown as an arrow with the label of the relationship name on each arrow.

4.2 Index Ontology

The index ontology represents knowledge between keywords and modifier in the ICD index. Relationships found in the index ontology include type, subtype, code as relationship (Fig. 3). During the experts’ work in building ICD-10-TM index ontology, it was found that the current ICD index structure which defines conditions complicating pregnancy and perinatal conditions at the same hierarchical level as organ system diseases (Fig. 4) could mislead a coding algorithm into a wrong selection of an ICD code. For example, under index keyword “hemorrhoids”, the modifier “complicating pregnancy” was in the same hierarchical level as the modifier “internal” and “external”. This could mislead a coding program into selecting an internal hemorrhoid branch instead of a complicating pregnancy branch in the case of pregnant women with an internal hemorrhoid, especially if a physician gave a diagnosis including the term “internal hemorrhoid”.

This flaw in the old index structure was found under keywords of most diseases affecting pregnancy and perinatal con-

![Fig. 7](https://example.com/image7.png)

**Fig. 7** Input screen of semi-automated ICD-10-TM morbidity coding system
4.3 Usability of the ICD-10-TM Ontology in Semi-automated ICD-10-TM Coding Software

The ICD-10-TM ontology was implemented in semi-automated ICD-10-TM coding software as a knowledge base (Fig. 6). The ontology provided results to the search algorithm of the system during index matching and tabular list verification processes.

ICD-10 is not an easy coding system. We cannot just try to match each diagnosis word with ICD-10 code directly. In ICD-10, any diagnosis word could be coded to various ICD codes, depending on type of patients (pregnancy, newborn, or general patient), etiology (trauma, non-trauma) or other conditions. The coding algorithm should imitate manual coding processes by searching through the index ontology and then verifying the code with the tabular list ontology.

As shown in Fig. 7, the coding software will get input from patient discharge summary forms which include age, sex, pregnancy, childbirth or postpartum conditions of the patient, all physician diagnosis phrases, etiology (trauma, congenital, neoplasm, infections) of each disease, location of each disease.

The coding algorithms search matching keywords and modifiers from the index ontology and diagnosis knowledge bases, then verify code definition and the include and exclude conditions from a tabular list ontology. The program will display all ICD-10-TM codes found or not found to the clinical coder as in Fig. 8. The human coder could accept the codes or change to other codes based on their judgment and standard coding guidelines.

The semi-automated ICD-10-TM coding system was released as a beta version on the Thai Health Coding Center website [35] on January 2011. Forty ICD coders from different hospitals volunteered to download, install and test the system during January to March 2011. The preliminary user report (Table 2) showed that users are satisfied with searching The ICD-10-TM ontology and thought that the ontology could be used as a knowledge base in semi-automated coding systems as well as in ICD-10-TM manual coding books.

5. Discussion

Despite recent published work [36] mentioning ICD-10 ontology creation, our work is the first ICD-10 ontology with a comprehensive description of all concepts and relations in ICD-10 tabular list and alphabetical index. A complete ICD-10 ontology should contain two major parts, i.e., tabular list and index ontology.

The ICD-10 alphabetical index book contains some implicit information about types of patient which will affect ICD code selection in the lead term searching process. Human coders, fully aware about the

<table>
<thead>
<tr>
<th>Items</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of users</td>
<td>40</td>
</tr>
<tr>
<td>Overall satisfaction of system</td>
<td>4.15 ± 0.34</td>
</tr>
<tr>
<td>Percentage of results got from</td>
<td>96.17 ± 2.64</td>
</tr>
<tr>
<td>index ontology search</td>
<td></td>
</tr>
<tr>
<td>Percentage of results got from</td>
<td>98.35 ± 1.01</td>
</tr>
<tr>
<td>tabular list ontology search</td>
<td></td>
</tr>
<tr>
<td>Opinion whether The ICD-10-TM</td>
<td>All Yes</td>
</tr>
<tr>
<td>ontology could be used instead of</td>
<td></td>
</tr>
<tr>
<td>books?</td>
<td></td>
</tr>
</tbody>
</table>

Table 2 Users report on the ICD-10-TM ontology evaluation compared with ICD-10-TM books
type of patients while doing the work, would easily choose the right context for each patient. However, for a semi-automated or automated coding system, index structures must be modified to match a search algorithm which will never know about the type of patients (implicit knowledge). Adding a type of patient field into the input screen of the system, could also help the search algorithm to correctly select the right code in index ontology.

ICD-10 coding is not a word matching process. Qualified human ICD coders will never do simple diagnosis word search or browse for the diagnosis term from a list of ICD codes and labels. Unfortunately, research on semi-automated and automated ICD coding system in the past [30–33] has not recognized this important concept. We would like to suggest that in the future all researchers interested in the development of ICD coding systems should be aware of the complexity of ICD-10 knowledge bases and coding processes before attempting to build their systems. Given by the fact that physicians will have more opportunity to use the system to encode their discharge diagnosis [37], these physicians will need robust system with high accuracy coding algorithm to ensure correct ICD encoding.

Despite The ICD-10-TM ontology being used in this semi-automated ICD coding system, the result of the study would not change if we use the ontology created from WHO ICD-10 because the Thai modification of ICD did not alter the main structure of the tabular list and classification system. The standard coding process was also endorsed for coding ICD-10-TM in Thailand. The only one difference using WHO ICD-10 ontology would be the inability to find the TM codes.

At this moment, ICD-10-TM index ontology will differ from printed version of ICD-10-TM by raising conditions complicating pregnancy and perinatal conditions into an upper hierarchical level of index. This change will be found in next printed version of ICD-10-TM index. This issue of index structure flaw was also reported to WHO in the annual ICD development meeting in October, 2010 [38]. For user of the Thai semi-automated coding system, if a patient type was specified as pregnancy or newborn, the searching algorithm will search in the index ontology accordingly. After searching, the system will provide tabular list of the ICD codes as well as inclusion and exclusion terms to the user and help s/he to complete next coding steps.

For future development of The ICD coding algorithm like in ICD-11 development, the underpinning structure of ICD described in tabular list ontology should be included into manual coding process as well as automated process. Although this issue was implicit mention in coding process [vii] (be guided by any inclusion or exclusion terms under the selected code or under the chapter, block or category heading). It would be better to tell the coder to check the core structure of the ICD chapters based on patient and disease type.

Medical experts in the research team also suggested building a 3rd knowledge base based on diagnosis phrases used by physicians, which could be called a Diagnosis Knowledge Base and could be used as a supplementary knowledge base. The obvious case for the use of this extra knowledge base is to provide a help feature when a user of the program begins typing some characters in the diagnosis word. The program would then show all available diagnosis words to help the user complete the search term without the need to fully type in all characters. This help feature can be seen in today search engines, e.g., in Google search.

The development of a semi-automated ICD coding system in Thailand is now (2011) continuing. At present, we are trying to make the system more automated, include a Thai language search module as well as the ICD-10-TM procedures coding support into the system.

6. Conclusion

The requirement of a knowledge base for ICD-10-TM morbidity coding software was satisfied with an ontology driven design. Processing of application models described with the help of ontological models allowed coding of a complex system like ICD feasible. Finally, ICD-10 ontology can be used as a basic infrastructure for building the WHO family of international classification (FIC) ICD ontology and the ICD coding software. This ontology and software can be used to enabling successful implementation of ICD in developing countries, especially in countries without an adequate number of competent ICD coders.

Acknowledgments

This research was supported by the Thai National Health Security Office, Thai Health Standard Coding Center (THCC), Ministry of Public Health, Thailand and Thai Collaborating Center for WHO-Family of International Classification. The author would like to thank Dr. Elvin James Moore for manuscripts reviewing and suggestion and thank anonymous reviewers for helpful and useful comments and suggestions.

Reference

S. Nitsuwat, W. Paoin: ICD-10-TM Ontology for a Semi-automated ICD Coding System

Methods Inf Med 6/2012 © Schattauer 2012


