Original Article

Healthc Inform Res. 2010 June;16(2):82-88.
doi: 10.4258/hir.2010.16.2.82
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Evaluation of the Clinical Data Dictionary (CiDD)

Myung Kyung Lee, MPH[1], Hyeoun-Ae Park, PhD[1], Yul Ha Min, MS[1], Younglan Kim, MS[1], Hyo Ki Min, BS[2], Sung Woo Ham, BS[2]

[1]College of Nursing, Seoul National University; [2]Korea Institute of Radiological & Medical Sciences, Seoul, Korea

Objectives: The purpose of the study was to evaluate content coverage and data quality of the Clinical Data Dictionary (CiDD) developed by the Center for Interoperable EHR (CiEHR). **Methods:** A total of 12,994 terms were collected from 98 clinical forms of a tertiary cancer center hospital with 500 beds. After data cleaning, 9,418 terms were mapped with the data items of the CiDD by the research team, and validated by 30 doctors and nurses at the research hospital. **Results:** Mapping results were classified into five categories: lexically mapped; semantically mapped; mapped to either a broader term or a narrower term; mapped to more than one term and not mapped. In terms of coverage, out of 9,418 terms, 6,750 (71.7%) terms were mapped; 4,319 (45.9%) terms were lexically mapped; 2,431 (25.8%) were semantically mapped; 281 (3.0%) terms were mapped to a broader term; 43 (0.5%) were mapped to a narrower term; and 550 (5.8%) were mapped to more than one term. In terms of data quality, the CiDD has problems such as errors in concept namingand representation, redundancy in synonyms, inadequate synonyms, and ambiguity in meaning. **Conclusions:** Although the CiDD has terms covering 72% of local clinical terms, the CiDD can be improved by cleaning up errors and redundancies, adding textual definitions or use cases of the concept, and arranging the concepts in a hierarchy.

Keywords: Clinical Terminology, Data Dictionary, Semantic Interoperability, Standard Terms

Received for review: May 7, 2010
Accepted for publication: June 22, 2010

Corresponding Author
Hyeoun-Ae Park, PhD
College of Nursing, Seoul National University, 28 Yeongeon-dong, Jongno-gu, Seoul 110-799, Korea. Tel: +82-2-740-8803, Fax: +82-2-765-4103, E-mail: hapark@snu.ac.kr

I. Introduction

Standardization in the field of health information becomes important as computer-based information systems and electronic health records are being rapidly introduced into health care sectors around the world. Standardization of clinical terminology is the foundation of a clinical information system and the central building block that supports communication across the different clinical information systems and achieves semantic interoperability [1].

Unfortunately, most health care application packages and institution-based health information systems have their own terminologies, resulting in overlooked synonymy and semantic collisions among concepts, which in turn produce non-interoperable patient data. Furthermore, most countries have designated more than one health care terminology and classification standard for electronicmedical records instead of recommending one single terminology and classification. For example, the United States, United Kingdom, Canada, and Australia recognize more than one health care terminology and classification [2]. A solution for this problem is to make a data dictionary to link different terminologies and classifications.

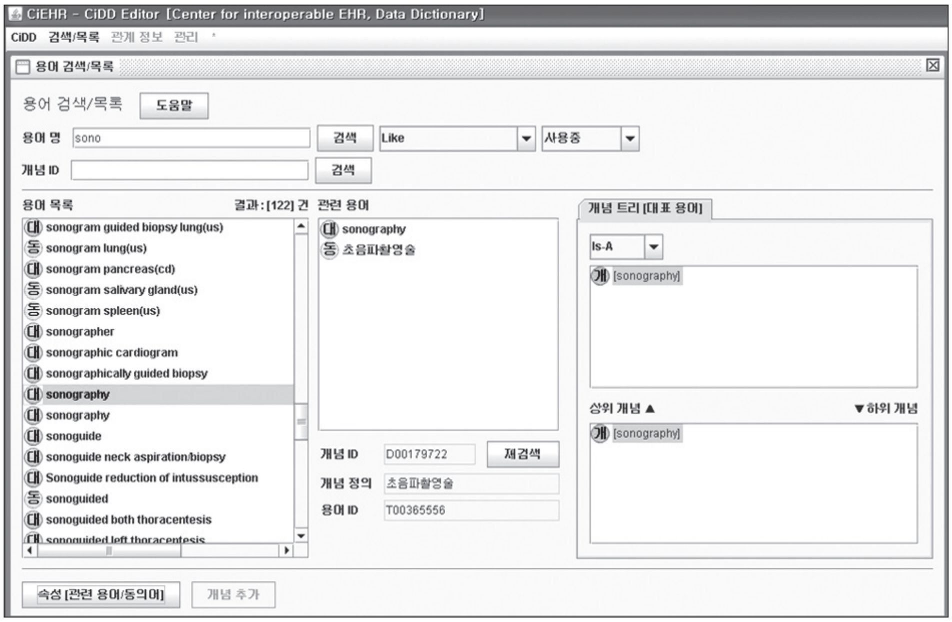
The need for a data dictionary becomes apparent when given the multiple definitions of a single term by different users not only acrossdifferent health care organizations but also within the same health care organization. Without a data dictionary, it is often difficult to build accurate and consistent patient records that can be shared across health care organizations. The Clinical Data Dictionary (CiDD) was developed by the Center for interoperable EHR (CiEHR) as a centralized repositoryof information about data such as names, meanings, types, formats, ranges of values, sources, and relationships to their data for each data element to represent a semantic relationship between data elements used in local hospital information systems and the standardized terminology of medicine developed in Korean Standard Terminology of Medicine. With the CiDD, local terminologies in health care institutions can be mapped with standardized terminologies and classifications for data sharing and exchange. This data dictionary was developed with an expectation of being used in hospital information systems throughout Korea.

The CiDD contains162,050 concepts and more than 427,276 terms with definitions and 155 value sets covering disease, clinical findings, and procedures. The concepts and terms of the CiDD were mapped to the KOSTOM, the KCD5 (Korean Classification of Disease, 5th Revision), the ICD9CM (International Classification of Diseases, 9th Revision, Clinical Modification), and the SNOMED-CT (Systematized Nomenclature of Medicine-Clinical Terms). A major part of the concepts, 160,888 concepts in total, are from the KOSTOM developed by the National Health Information Task Force Team. The CiDD was designed to provide a common health care language for clinical data to be indexed, stored, retrieved, and aggregated across specialties and sites of health care. It is designed for use in electronic medical records, reducing variability in the way data are captured, encoded, and used for the clinical care of patients and research.

This study was proposed to test if the CiDD can be used in hospital information system development. The research questions were first, to what extent can the CiDD cover the content of local terminology? Second, how good is the quality of data items in the CiDD? To answer these questions, we mapped the local terms from a tertiary cancer hospital with 500 beds to the data items of the CiDD. With this study, we hope to contribute to the refinement of the local vocabularies by providing preferred terms used in the CiDD. We also hope to contribute to the improvement of the CiDD by proposing new data items and providing ideas for the improvement of the data quality of the CiDD.

II. Methods

This study was conducted from May 2009 to January 2010. The research hospitalwas a tertiary cancer hospital with 500 beds. The hospital currently has an order communication system (OCS), and a picture archiving and communication system (PACS). The enterprise EMR system is planned to be introduced in January 2010. Local terms with mapping information to the CiDD will be used in the enterprise EMR



용어 명 sono

Like

사용중

개념 ID

용어 목록 결과 : [122] 건

- 대 sonogram guided biopsy lung(us)
- 동 sonogram lung(us)
- 대 sonogram pancreas(cd)
- 동 sonogram salivary gland(us)
- 동 sonogram spleen(us)
- 대 sonographer
- 대 sonographic cardiogram
- 대 sonographically guided biopsy
- 대 sonography
- 대 sonography
- 대 sonoguide
- 대 sonoguide neck aspiration/biopsy
- 대 Sonoguide reduction of intussusception
- 동 sonoguided
- 대 sonoguided both thoracentesis
- 대 sonoguided left thoracentesis

관련 용어

- 대 sonography
- 동 초음파촬영술

개념 ID D00179722

개념 정의 초음파촬영술

용어 ID T00365556

개념 트리 [대표 용어]

Is-A

개 [sonography]

상위 개념 ▲ ▼ 하위 개념

개 [sonography]

Figure 1. Example of a web-based Clinical Data Dictionary editor for searching terms.

system. The study consisted of four phases: collecting local terms; extracting unique concepts; mapping the terms and concepts; and validating the mapping.

1. Collecting Local Terms

To collect clinical terms from the research hospital, we first collected all the clinical forms used in the hospital. We extracted local terms from 98 clinical forms which included 22 forms used by nurses, 71 forms used by doctors, and 5 forms used by other health care professionals. We cleaned up the extracted terms by translating abbreviated terms into full names, correcting typos, and unifying capital and lowercase English letters.

2. Extracting Unique Concepts

We extracted unique concepts from the local terms by removing redundant and synonymous terms. If a term was a compound concept that consisted of more than two concepts, we divided the compound terms into atomic concepts based on the units of meaning.

3. Mapping

The research team consisting of 5 nursing informatists mapped local terms to data items of the CiDD using the CiDD editor (Figure 1) [3]. The CiDD editor is a tool for searching a term in the CiDD. Figure 1 shows a sample screenshot of the CiDD editor showing the data-search window and the display window of searched terms. The left upper part of the screen shows the blank box where a term can be input for search. There are "Like", "Start with and like", "Match against", "Start with", "End with", and "Exact" searching options available. The lower part of the screen displays the list of searched terms. When one term is selected out of these searched terms, the upper middle part of the screen shows the representative term and synonyms. The lower part shows the concept ID, the definition of the concept, and the term ID.

We created a mapping table containing data source information, target information, and mapping results. Data source information includes which section of a form of a department a term is from. Target information includes whether it is a preferred term or a synonym with a concept ID and term ID. Mapping results includes whether it is mapped or not mapped, along with a new term proposed in case it is not mapped.

The detailed mapping process is described in Figure 2. The first phase of matching is linguistic matching based on term labels. Label matching involves putting the label into a canonical form by stemming and tokenization; comparing the equality of labels; and matching sub-strings [4]. Term names with suffixes such as verb variations (ex: assessing vs. assessment vs. assesses) and singular versus plural words (medication vs. medications, site vs. sites), uses of preposition (ex, monitoring vs. monitoring for, screening vs. screening for, implementation vs. implementation of), compound words with or without spaces (well being vs. wellbeing), and

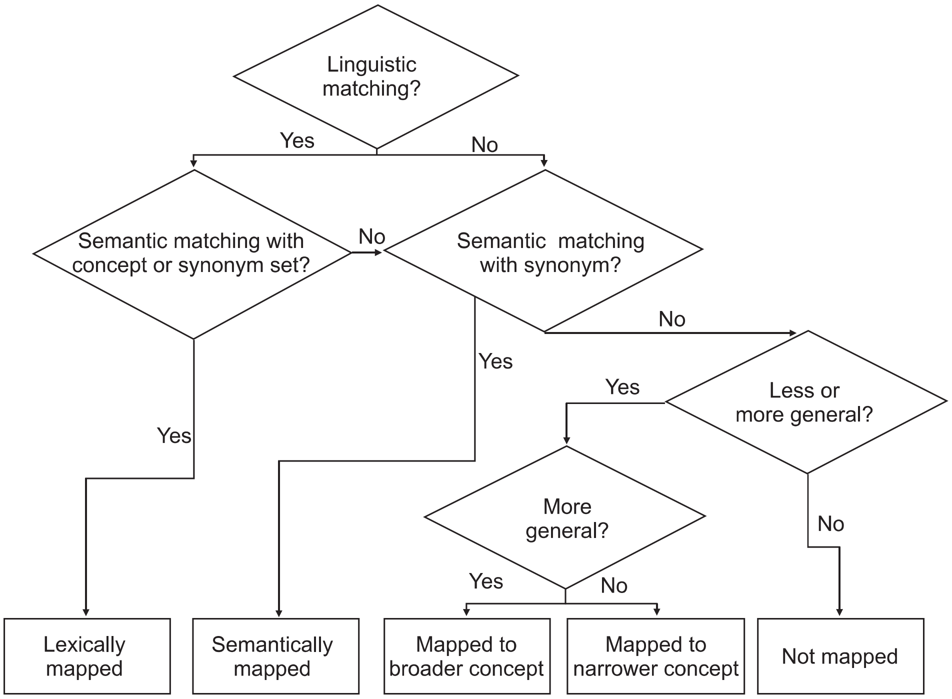

Linguistic matching?
Yes
No
Semantic matching with concept or synonym set?
No
Semantic matching with synonym?
No
Yes
Yes
Less or more general?
Yes
More general?
No
Yes
No
Lexically mapped
Semantically mapped
Mapped to broader concept
Mapped to narrower concept
Not mapped

Figure 2. Mapping process of local terms to Clinical Data Dictionary terms.

doi: 10.4258/hir.2010.16.2.82

compound words with or without hyphens (ex: self toileting vs. self-toileting) were treated as linguistically identical. If a term in local terms matched linguistically with a term in the CiDD, the next phase was the semantic matching with its concept or synonym set based on the similarities of their contexts or vicinities in the terms and concepts. We used the definition of the local term and synonym set as our first mapping criteria. If a term in local terms matched with a term in the CiDD linguistically and semantically, then we classified this as completely mapped (lexically mapped). If a term in local terms did match linguistically but not semantically with a term in the CiDD, we examined if it was semantically matched with another concept in the CiDD. Semantic matching is an approach where semantic relations are examined between terms (not between labels) based on definitions [5,6]. For semantic matching, we came up with other expressed terms with the same concept and searched these synonyms for semantic mapping using the multiple searching options of the CiDD editor. If a term in local terms matched with a term in the CiDD semantically, then we classified this as semantically mapped. If a local term matched to a more general CiDD term, it was classified as mapped to a broader term. If a local term matched to a less general term in the CiDD, it was classified as mapped to a narrower term. If a local term mapped to more than one term, it was classified as mapped to more than one term. Otherwise, it was classified as not mapped.

4. Validation

The mapping was validated by 30 nurses and doctors who are employed by the research hospital and are users of the clinical terms. We explained the mapping process and results to the validation team. We then examined each concept with the validation team. During the validation process, we asked the validation team whether we translated abbreviated terms into full terms correctly, whether we divided compound concepts based on units of meaning correctly, whether we understood the meaning of concepts correctly, and whether we mapped correctly. If there was any disagreement between the initial mapping team and the validation team, we convened another meeting to reach a consensus.

III. Results

1. Collecting Terms and Cleaning

We collected 12,994 local terms from the 98 clinical forms used in 23 departments of the research hospital. We translated abbreviated terms into full terms, unified all English terms by using lowercase letters only except in proper nouns, and corrected typographical errors. After the cleaning process, 9,640 terms were identified. Table 1 shows the number of terms and clinical forms by department. The terms used in the medical departments accounted for 72% of the total terms, and the nursing department, 20% of the total local terms.

Table 1. Number of clinical forms used and local terms collected by departments

Department	No. of clinical forms	No. of local terms (%)
Nursing	22	1,966 (20.4)
Internal medicine	4	254 (2.6)
Anesthesia	2	179 (1.9)
Radiology	4	631 (6.5)
Urology	7	501 (5.2)
Social work	3	144 (1.5)
Obstetrics & Gynecology	5	439 (4.6)
Pediatrics	2	196 (2.0)
Neurology	20	850 (8.8)
Psychiatrics	1	68 (0.7)
Ophthalmology	1	42 (0.4)
General surgery	14	1,753 (18.2)
Emergency medicine	7	557 (5.8)
Otorhinolaryngology	10	660 (6.8)
Medical examination	1	100 (1.0)
Common forms for all	7	307 (3.2)
Orthopedics	4	117 (1.2)
Dental	1	51 (0.5)
Dermatology	1	23 (0.2)
Nuclear medicine	2	75 (0.8)
Thoracic surgery	3	226 (2.3)
Cyber knife center	2	205 (2.1)
Nutrition	2	296 (3.1)
Total	125	9,640 (100)

2. Mappings

Table 2 shows the initial mapping result. Out of 9,640 terms, 4,486 (46.5%) terms were lexically mapped, and 2,411(25.0%) were semantically mapped. 244 (2.5%) terms were mapped to a broader term, 48 (0.5%) terms were mapped to a narrower term, 583 (6.0%) terms were mapped to more than one less general term, and 1,467 (15.2%) terms were not mapped. There were 401 (4.2%) terms that we could not map because we did not understand the meaning of terms fully.

Table 2. Mapping results of local terms and unique concepts to the Clinical Data Dictionary before and after validation

Mapping result	No. of local terms (%)		No of unique concepts (%)	
	Before validation	After validation	Before validation	After validation
Lexical mapping	4,486 (46.5)	4,319 (45.9)	1,404 (35.6)	1,361 (37.4)
Semantic mapping	2,411 (25.0)	2,431 (25.8)	840 (21.3)	807 (22.2)
Broader term mapping	244 (2.5)	281 (3.0)	136 (3.44)	144 (4.0)
Narrower term mapping	48 (0.5)	43 (0.5)	24 (0.6)	19 (0.5)
One to many terms mapping	583 (6.0)	550 (5.8)	419 (10.6)	351 (9.6)
Not mapped	1,467 (15.2)	1,794 (19.0)	901 (22.9)	959 (26.3)
Unable to do mapping (need to find full term or meaning)	401 (4.2)	0 (0.0)	219 (5.6)	0 (0.0)
Total	9,640	9,418	3,943	3,641

Out of 9,640 local terms, 3,943 unique concepts were extracted by removing redundancy and synonyms. Out of 3,943 unique concepts, 1,404 (35.6%) concepts were lexically mapped, and 840 (21.3%) concepts were semantically mapped. 136 (3.44%) concepts were mapped to a broader term, 24 (0.6%) concepts were mapped to a narrower term, 419 (10.6%) concepts were mapped to more than one concept, and 901 (22.9%) concepts were not mapped. There were 219 (5.6%) concepts that we could not map to the CiDD due to ambiguity.

3. Validation

During the validation process, with 30 clinicians from the research hospital, we validated the initial mapping results and clarified the meaning of ambiguous terms that we could not map during the initial mapping process. 184 mappings were found to be invalid. For example, we mapped 'subtitle' from the operation record of the neurology department to 'subdiagnosis'. However, it was found that 'subtitle' is used for describing 'tumor site' or 'tumor location'. Another example is 'Imaging' from the short-term admission record of the general surgery department, which we mapped to 'image'. However, it was found that 'Imaging' is used to describe an imaging 'diagnostic test'. We mapped 'fungus' from the dermatology outpatient's initial record to 'fungus'. It was found that it is used to describe the 'allergy test for funguses'. We classified 'absorbed dose' from the radiology department into not mapped, however, we found that it is used to describe the 'accumulated dose' of radiology and we changed the mapping result to lexically mapped.

In addition through the validation process, we were able to identify 222 terms not used any longer. As a result, 9,418 terms remained. The lexically mapped category decreased from 46.5% (in unique concepts, 35.6%) to 45.9% (in unique concepts, 37.4%). The semantically mapped category increased from 25.0% (in unique concepts, 21.3%) to 25.8% (in unique concepts, 22.2%). The mapped to a broader term category increased from 2.5% (in unique concepts, 3.44%) to 3.0% (in unique concepts, 4.0%), and the mapped to a narrower term category remained the same at 0.5% (in unique concepts, 0.6% to 0.5%). The mapped to more than one term category decreased from 6.0% (in unique concepts, 10.6%) to 5.8% (in unique concepts, 9.6%). The not mapped category increased from 15.2% (in unique concepts, 22.9%) to 19.0% (in unique concepts, 26.3%).

IV. Discussion

This study was conducted to evaluate the content coverage and data quality of the Clinical Data Dictionary (CiDD). We collected local terms from a tertiary cancer hospital with 500 beds and mapped local terms to data items of the CiDD. In this section, we would like to discuss problems we encountered during this study and ways to improve the CiDD.

We encountered many problems when we collected local terms from the research hospital. First of all, there was no terminology specialist in the research hospital we could refer to when we encountered any problems during the collection of local terms. We also found that many local terms were used inappropriately with spelling errors, incorrect transcriptions in English, and locally specified abbreviations. Also, there were many redundant expressionswith different variations and different value sets. Thus, a considerable amount of effort was required to refine the local terms and extract core conceptsfrom these terms before mapping.

We found that 71.7% of terms and 56.8% of unique concepts collected from the research hospital were mapped to the CiDD semantically. Most terms or concepts not mapped to the CiDD were terms or concepts describing nursing assessment, nursing education, or procedures such as surgery and treatment. This could be explained by the history and scope of the KOSTOM, which is the major component of the CiDD. The KOSTOM was developed to document medical diagnoses, laboratories, treatments, and anatomical sites in the beginning. Thus, in order for the CiDD to be used for hospital information systems, terms describing other domains of health care need to be added. We suggested adding local terms not mapped, mapped to either a broader or a narrower term, or mapped to more than one term as new terms to the CiDD. We also suggested adding more synonyms to the CiDD to describe semantically mapped local terms.

Validation was performed both internally and externally. First, mapping results were compared across research members internally, and differences were resolved by a consensus process, and then validated externally by doctors and nurses who are the users of the clinical terms. There was little differencein mapping before and after users' validation. This might be due to the fact that the ambiguity of terms or concepts was resolved during the mapping process by consulting medical record administrators or users of the research hospital. Most of the validation time was spent clarifying 401 ambiguous terms and uncertain abbreviations.

Through this study we found that the CiDD was able to represent most local terms. However, we also found that the CiDD has several data quality issues. First, data items in the CiDD have an accuracy or validity [7-9] problem with definition. The definition of most data items in the CiDD is the same as the representative term or concept name, without textual definition or formal definition. There are multiple terms or concepts with the same definition. For example, 'Alleviator', 'Demulcents', 'Emollients', 'Lenitive', 'Malactic', 'Malagma', and 'Torpent' had the same definition, "Wanhwaje", which means something for relaxing, easing, relieving, alleviating, softening, lightening and mitigating.

There are mismatches between concepts and representative terms, and mismatches between concepts and synonyms. For example, the concept 'divorce [D00150364]' and the representative term 'marital separation [T00345013]' are a mismatch, and the representative term 'current drinker [T00381973]' and the synonym 'alcohol [T00427132]' are a mismatch.

Second, data items of the CiDD have a completeness [7-9] problem. In the CiDD, there is a 'Healthy looking Appearance [T00426281]', 'Acute ill looking Appearance [T00426277]', and 'Chronic ill looking Appearance [T00426278]', but not an 'Ill looking appearance'. 'Chronic ill looking appearance' and 'Acute ill looking appearance' should be children concepts of 'Ill looking appearance'. 'Healthy looking appearance' and 'Ill looking appearance' should be children concepts of 'Looking appearance'.

Third, data items of the CiDD have a redundancy [7-9] problem in both representative terms and synonyms. An example of redundancy in representative terms with the same definitions is 'Gynecological examination [T00299305]' and 'Gynecological examination [T00299303]'. An example of redundancy in representative terms with different definitions is 'lithotomy position [T00277785]' with the definition Doljegeosool Jase', which means a body position for surgical procedures to remove a calculus from the kidney, and 'Lithotomy position [T00277787]' with the definition "Doljegeo Jase" which means a body position to remove a calculus from the kidney. Examples of redundancy due to different ways of representing terms or concepts such as with or without hyphens '-', with or without spaces between words, and using capital or lowercase letters are 'serum glutamic oxaloacetic transaminase [T00425157]' and 'Serum glutamic-oxaloacetic transaminase [T00383881]', and 'Anterior [T00327085]' and 'anterior [T00353034]'. In terms of the redundancy of synonyms, examples with identical synonyms are 'BUN [T00199721]' and 'BUN [T00383428]', and 'IBW [T00145452]' and 'IBW [T00343643]'. Example of redundancy in synonyms due to different ways of representing are as follows: an example with spaces between words is 'Hyeolaeg Yoso Jilso [T00199717]' and 'Hyeolaegyosojilso [T00199718]' ; an example of singular versus plural expressions is 'upper and lower extremity [T00374699]' and 'upper and lower extremities [T00374698]' an example with or without hyphens is 'computer assisted tomography [T00353074]' and 'computer-assisted tomography [T00353075]' an example with or without '.' in an abbreviation is 'Hb [T00383124]' and 'Hb. [T00383125]'.

Fourth, data items of the CiDD had a comprehensiveness or coverage [8,9] problem. The CiDD lacks clinical terms frequently used in clinical practice. For example, there was 'Mu' which means 'not present', or 'no', but there was no concept or term to describe 'Yu' which means 'present', or 'yes'.

Fifth, data items of the CiDD had a currency [8,9] problem. Terms related to current medical science such as cyber knife or robot surgery were not reflected sufficiently in the CiDD. Examples are 'Image Guidances', and 'RoboCouch'.

Sixth, data items of the CiDD had a consistency [8,9] problem in describing concepts. For example, the concept 'Brother [T00384176]' has a synonym, 'brothers [T00384177]'. However, the term 'sisters [T00349544]' has no 'sister' term as a synonym in the CiDD, which is more commonly used than

'Sisters'.

We recommend that the data quality issues of the CiDD listed above be resolved before it is distributed widely as a standardized data dictionary in Korea. Also, we recommend adding textual or context-sensitive definitions, use cases of the concept or term, value sets, or hierarchies to avoid the possibility of concepts or terms being interpreted differently by different users.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgements

This study was supported by a grant of the Korea Healthcare Technology R&D Project, Ministry for Health, Welfare & Family Affairs, Republic of Korea (A050909).

References

1. AHIMA e-HIM Workgroup on EHR Data Content. Guidelines for developing a data dictionary. J AHIMA 2006; 77: 64A-64D.
2. American Medical Informatics Association and American Health Information Management Association Terminology and Classification Policy Task Force. Healthcare terminologies and classifications: an action agenda for the United States. Bethesa (MD): American Medical Informatics Association; 2006 [cited at 2010 June 24]. Available from: http://library.ahima.org/xpedio/groups/public/documents/ahima/bok1_032395.pdf.
3. Yun JH, Kim MJ, Ahn SJ, Kwak MS, Kim Y, Kim HK. The development of clinical terminology dictionary for integration and management of clinical terminologies in EMR Systems. J Korean Soc Med Inform 2009; 15: 411-421.
4. Wang Y, Patrick J, Miller G, O'Halloran J, eds. Linguistic mapping of terminologies to SNOMED CT. Proceedings of Semantic Mining Conference on SNOMED-CT; 2006 Oct 13; Copenhagen. [cited at 2010 Mar 3]. Available from: http://jodi.tamu.edu/Articles/v01/i08/Doerr/.
5. Avesani P, Giunchiglia F, Yatskevich M. A large scale taxonomy mapping evaluation. Lect Notes Comput Sci 2005; 3729: 67-81.
6. Doerr M. Semantic problems of thesaurus mapping. J Digit Inf [Internet]. 2001 [citied at 2010 Mar 3]; 1. Available from: http://journals.tdl.org/jodi/article/view/31.
7. Choe IS. Evaluation and quality control of data in the digital library system. J Korean Soc Libr Inf Sci 2004; 38: 119-139.
8. Yeganeh NK, Sadiq S, Deng K, Zhou X. Data quality aware queries in collaborative information systems. In: Li Q, Feng L, Pei J, eds. Advances in data and web management. Berlin: Springer; 2009. p39-50.
9. Pipino L, Lee YW, Wang RY. Data quality assessment. Commun ACM 2002; 45: 211-218.