ICBO: International Conference on Biomedical Ontology
July 28-30, 2011 · Buffalo, NY, USA

A Case Study of ICD-11 Anatomy Value Set Extraction from SNOMED CT

Guoqian Jiang[1], Harold R. Solbrig[1], Robert J.G. Chalmers[2],
Kent Spackman[3], Alan L. Rector[2], Christopher G. Chute[1]

[1]Mayo Clinic College of Medicine, Rochester, MN. U.S.A.
[2]University of Manchester, Manchester, U.K.
[3]International Health Terminology Standards Development Organisation, Copenhagen, Denmark

Abstract. The 11th revision of the International Classification of Diseases (ICD-11) intends to derive its "ontological component" from external ontologies (e.g. SNOMED CT). One of the core value sets is an ICD-11 anatomy chapter. The objective of the present study is to develop and evaluate approaches to value set extraction from SNOMED CT for the ICD-11 anatomy use case. We investigated a number of resources comprising SNOMED CT base terms, the anatomical term mappings between ICD-O (ICD for Oncology) and SNOMED CT, the CORE Problem List Subset of SNOMED CT, and the SNOMED CT stated form in Web Ontology Language (OWL). We used the Manchester OWL module extraction tool and its extension in Protégé 4.1. We proposed and evaluated four semi-automatic value set extraction strategies based on different clinical contexts and discussed their implications in terms of domain coverage, granularity and clinical usefulness from both technical and clinical perspectives.

Keywords: ICD-11, SNOMED CT, Web Ontology Language (OWL), Ontology modularity, Value set Definition

1 Introduction

The 11th revision of the International Classification of Diseases (ICD) was officially launched by the World Health Organization (WHO) in March 2007 [1]. The WHO has sought to reuse existing ontologies such as SNOMED CT for value set definition. One of the core value sets being developed is for anatomical site, defined by WHO as "the most specific level of the topographical location or the anatomical structure where the health-related problem can be found relevant to the condition" [2].

In this context, a value set is a uniquely identifiable set of valid values that can be associated with a defined set of ICD entities. Typically, value sets can be drawn from pre-existing coding schemes such as SNOMED CT by constraining the value selection based on a logical expression (e.g. all sub-codes of the code "breast cancer"). Generating clinically meaningful value sets in a (semi-) automatic way from a terminology/ontology service has been challenging for the community, in part due to the lack of 1) formal linkage to clinical context patterns that act as constraints in defining a concept domain; 2) techniques for automatically linking values to their appropriate concept domains, and 3) tools based on formal language such as the Web Ontology Language (OWL) [3]. To deal with some of these challenges, a number of research and standardization efforts are being undertaken, including HL7 Common Terminology Services II specification [4], Mayo's LexEVS 6.0 implementation on a value set definition service [5], and Manchester's new OWL API 3 [6] which contains a set of modularization tools [7].

In the present work, we performed a case study of ICD-11 anatomy value set extraction from SNOMED CT. We propose four semi-automatic value set extraction strategies based on different clinical context patterns. We evaluated the strategies and discuss their implications in terms of domain coverage, granularity and clinical usefulness from both technical and clinical perspectives.

2 Background

2.1 ICD-11 and its Content Model

WHO has embraced a broadened set of use cases to drive ICD-11 development [8]. The purpose of the ICD-11 content model is to present the knowledge that underlies the definitions of an ICD entity [2]. Table 1 illustrates that "Body System/Structure Description" is one of 13 main parameters for describing an ICD category.

1	ICD Entity Title
2	Classification Properties
3	Textual Definitions
4	Terms
5	Body System/Structure Description
6	Temporal Properties
7	Severity of Subtype Properties
8	Manifestation Properties
9	Causal Properties
10	Functioning Properties
11	Specific Condition Properties
12	Treatment Properties
13	Diagnostic Criteria

Table 1. The ICD-11 content model main parameters

2.2 SNOMED CT and its Concept Model

SNOMED CT is the most comprehensive clinically oriented medical terminology system. It is owned and maintained by the International Health Terminology Standard Development Organization (IHTSDO) [9]. The IHTSDO and the WHO signed a collaborative agreement in July 2010, which essentially establishes SNOMED CT as the core of the ontological component of ICD [10].

For its representation of anatomy, SNOMED CT has adopted a variant of the "Structure-Entire-Parts (SEP)" triple mechanism developed by Hahn and Schulz [11-12]. Fig. 1 shows a representation example for "skin structure of face", in which "skin structure of face", "entire skin of face" and "skin of part of face" forms a SEP triple.

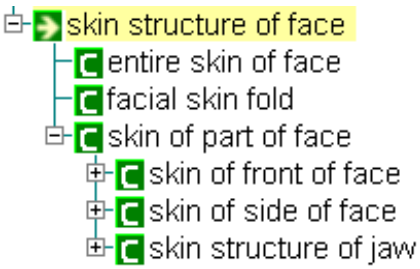

Figure 1. An example of SEP triple representation for "73897004 skin structure of face".

SNOMED CT has a Clinical Finding concept model, in which a set of attributes has been specified to define Clinical Finding concepts [13]. Two of these attributes, "Finding Site" and "Associated Morphology," are allowed values taken from the hierarchy under "123037004 Body structure". Table 2 shows the two anatomy-related defining attributes.

Defining Attribute	Subsumed Attribute	Allowable Values
|FINDING SITE|		|Anatomical of acquired body| 442083009 (<<)
|ASSOCIATED MORPHOLOGY|		|Morphologically abnormal structure| 49755003 (<<)

Table 2. Two anatomy related attributes specified in SNOMED CT Clinical Finding concept model

3 Materials and Methods

3.1 Materials

For this study we used:

1) The topographical term mappings from SNOMED CT to ICD-O Topography provided by the 20100731 International Release of SNOMED CT;

2) A subset of SNOMED CT anatomy "base terms" extracted by IHTSDO from its complete set of anatomical concepts [14];

3) A subset of "Clinical Finding" concepts extracted from SNOMED CT stated form after conversion into Web Ontology Language (OWL) using the Perl script provided by the original distribution;

4) The entries of the type "finding" or "disorder" extracted from a download of the UMLS CORE Problem List Subset of SNOMED CT [15]

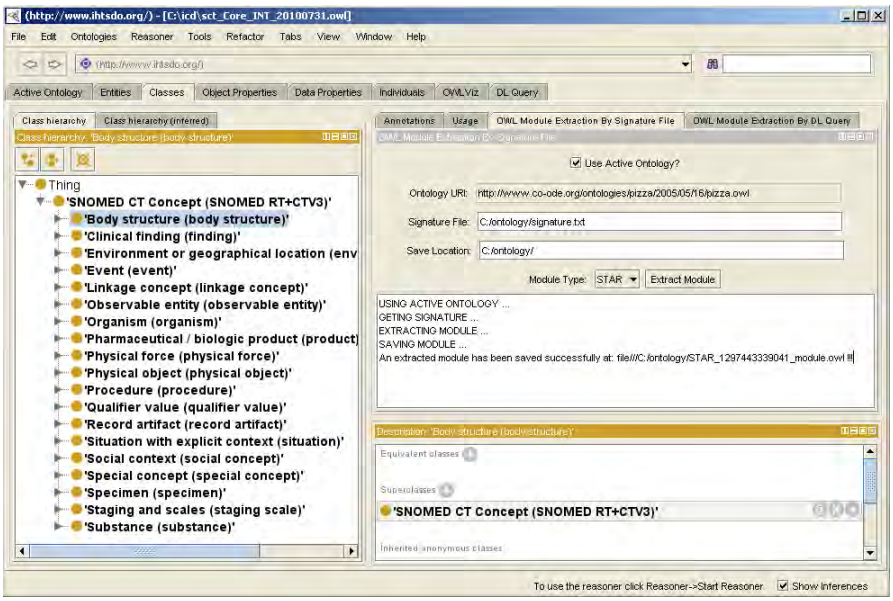

Figure 2. A screenshot of the SNOMED CT module extraction tool as a Protégé 4.1 plugin.

3.2 Methods

We took advantage of a Manchester SNOMED module extraction API [16, 17] and developed an extension as a Protégé 4.1 plugin [18]. In this way, we were able to load the SNOMED CT stated form as OWL into the Protégé 4.1 platform for module extraction. Fig. 2 shows a screenshot of the module extraction plugin in Protégé 4.1.

We defined four clinical context patterns for the ICD anatomy value set extraction, one for each of the sources described above. We created a signature file for each and identified concept IDs for 23,221 SNOMED CT concepts mapped to ICD-O topography, 14,871 concepts extracted from the anatomy base terms subset of SNOMED CT (note that the concept IDs are not available for a portion of base terms), 97,138 concept IDs extracted from the branch "Clinical Finding" of SNOMED CT and 5,304 concepts from the CORE Problem List subset of SNOMED CT.

With the four signature files, we generated four modules in OWL syntax using the module generation tool. For the first and second patterns, the generated modules are anatomy-specific modules, whereas for the third and fourth patterns, the generated modules are not, because the original signatures are all from the "Clinical Finding" domain. As the module extraction tool extracts all axioms relevant to the signatures, the corresponding anatomical structures are also extracted. Once the modules were generated in the first round, we created a signature file for each module using the concept IDs extracted from the "body structure" sub-tree of each module. Using the signature files, we generated the anatomy-specific modules for the third and fourth patterns.

We evaluated the anatomy-specific modules extracted from the four patterns in three aspects: 1) domain coverage, 2) module granularity, and 3) clinical usefulness of the module. For domain coverage, we used the 287 ICD-O topographical categories as anchors to classify the concept IDs in each module. The domain coverage is measured by the ratio of the number of categories containing mappings over total number of categories (i.e. the 287 categories). For module granularity, we defined two measures: general granularity and adjusted granularity. The general granularity is measured simply by the ratio of the number of concept IDs in the module over the number of concept IDs in a control module (i.e. the module of the first pattern). The adjusted granularity is measured by the average ratio of the number of concept IDs in each of 287 categories in a module over that of the control module.

In addition, we performed a preliminary evaluation of clinical usefulness of the modules. We chose two categories out of the 287 ICD-O

categories "C44.3 skin of face" and "C44.5 Skin of trunk" in the dermatology domain. One of the authors (RC, a dermatology physician) reviewed the SNOMED CT mapping concepts to the two categories and marked those that should be considered as part of ICD-11 anatomy concepts. We measured the clinical usefulness by the ratio of the number of concepts checked (by RC's ratings) over the number of total concepts in the two categories.

4 Results

In total, there are 31,107 concept IDs under the "123037004 Body structure" branch of SNOMED CT. We successfully extracted four anatomy-specific modules from SNOMED CT based on four different clinical context patterns. Table 3 shows the number of concept IDs in each module and their distribution. The majority of concept IDs in each module are those under "91723000 Anatomical structure", whereas in the modules *ClinicalFinding* and *ProblemList*, a significant number of concept IDs (1,919 and 755 respectively) are under "118956008 Morphologically altered structure".

Fig. 3 shows the evaluation results of domain coverage, general granularity and adjusted granularity of the four extracted modules. The results indicate that compared with the *Control* module, the *BaseTerm* module and *ClinicalFinding* module reduced granularity approximately by one third and two third respectively but still keep good domain coverage, whereas the *ProblemList* module reduced granularity by about nine tenths while it lost domain coverage by about one fourth.

Table 4 and Table 5 show the evaluation results of clinical usefulness (which is based on RC's ratings) for two target ICD-O categories. The results indicate that the *ClinicalFinding* and *ProblemList* modules had better outcome in terms of clinical usefulness, revealing that the clinical context patterns underlying the modules are effective and match more closely with the clinician's expectations on ICD-11 anatomy.

	Total	ANS	ACS*	MAS*	AOP	Qualifier
Control	24142	23617	63	73	14	4
BaseTerm	15167	15004	55	70	24	4
ClinicalFinding	7120	5218	46	1919	2	4
ProblemList	2955	2201	23	755	0	4

ANS – Number of Concept IDs under "91723000 Anatomical structure"

ACS – Number of Concept IDs under "280115004 Acquired body structure"

MAS – Number of Concept IDs under "118956008 Morphologically altered structure"

AOP – Number of Concept IDs under "91832008 Anatomical organizational pattern"

Table 3. Number of concept IDs in each module and their distribution
*Note that some of concept IDs under ACS and MAS overlap.

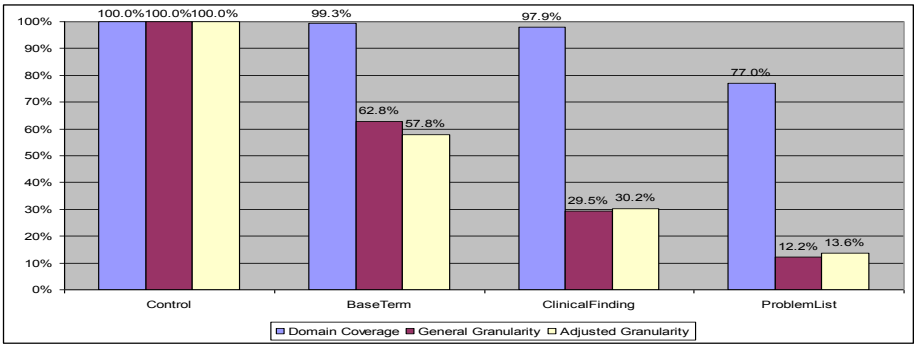

Figure 3. Evaluation results of domain coverage and granularity for four modules

	Number of Concepts Checked	Total Number	Ratio	Relative Ratio (vs. Control)
Control	17	70	24.3%	1.00
BaseTerm	17	48	35.4%	1.46
ClinicalFinding	14	34	41.2%	1.70
ProblemList	4	11	36.4%	1.50

Table 4. Evaluation results of clinical usefulness for category "C44.3 Skin of face". The number of concepts checked is based on RC's ratings.

	Number of Concepts Checked	Total Number	Ratio	Relative Ratio (vs. Control)
Control	11	149	7.4%	1.00
BaseTerm	11	85	12.9%	1.75
ClinicalFinding	11	56	19.6%	2.66
ProblemList	6	20	30.0%	4.06

Table 5. Evaluation results of clinical usefulness for category "C44.5 Skin of trunk". The number of concepts checked is based on RC's ratings.

5 Discussion

For the development of the ICD-11 anatomy chapter, the list of ICD-O topographical codes was a potential candidate [19] as it has been standardized in the oncology domain by WHO to describe the "Neoplasm" chapter of ICD-10. However, the list of ICD-O codes is sparse and not sufficiently detailed for many purposes.

As the IHTSDO provides mappings from SNOMED CT to ICD-O Topography as a part of the standard distribution, the mappings were naturally considered as a way to identify possible candidates for the purposes of ICD-11 anatomy. However, the problem with this idea is that the mapping is directional, and the direction is from SNOMED CT to ICD-O codes, meaning that virtually all SNOMED CT anatomy codes that could be relevant in cancer (i.e. that can be the site of a malignancy) are mapped. Moreover, the map doesn't identify a single best SNOMED CT code for each unique ICD-O topographical code.

The entire set of SNOMED CT codes included in the map does not make a significant reduction in the size of the anatomy terminology, and leaves us with the atypical and highly unfamiliar naming of things according to the S-E-P model, such as "X structure (body structure)" and "Entire X (body structure)". Based on the assumption that end users of the anatomy codes will want familiar names and a full set, IHTSDO created a subset of SNOMED CT which contains 18,266 base terms, from which we extracted 14,871 concept IDs for module extraction (note that some of the former have their origin with the FMA but have no corresponding SNOMED CT concept, and thus do not have a concept ID assigned).

For the SNOMED CT Clinical Finding concept model, we consider the attributes "Finding Site" and "Associated Morphology" to be analogous to the parameter "Body structure" in the ICD-11 content model, whereas the SNOMED CT concepts under the Clinical Finding branch are analogous to the disease categories in ICD-11, although the SNOMED CT concepts are more fine-grained. We consider that the asserted anatomical structures corresponding to the Clinical Finding concepts are meaningful to be a clinical context pattern for the ICD-11 anatomy value set.

We defined quantitative measures to evaluate the four modules in terms of domain coverage, module granularity and clinical usefulness. We believe that the measures are useful to help in deciding which module extraction strategy is effective and should be considered for adoption. Note that the usefulness evaluation we performed in the present study has limited generalizability because it is based on a single reviewer's ratings. It seems clear that we could obtain more reliable results by using more experts from diverse clinical backgrounds.

Based on our results, we suggest that the

strategy used for the module *ClinicalFinding* as a good starting point for the ICD-11 anatomy use case. The module has good domain coverage while keeping a relatively small size and better outcome on clinical usefulness. Note that the SNOMED CT anatomy base terms are still useful for providing familiar names and a full set, and are complementary to our suggested strategy here.

In addition, the upper level of the SNOMED CT anatomy may create some confusion because it has three main branches: 1) Anatomical structure, i.e. the normal anatomy, 2) Acquired body structure - mostly the results of surgery plus "scar (morphologic abnormality)", and 3) Body structure altered from its original anatomic structure (morphologic abnormality) – the results of disease or repair. Given these different types of body structure, the question really is what is needed for ICD-11 development. The clinical experts in the WHO TAG groups, editorial boards and the clinical groups working on anatomy will need to clarify their requirements in order to determine whether one or all of these branches should be used.

Finally, the Protégé based OWL module extraction tool we developed in this study [18] has been demonstrated very useful for achieving our goal. While we mainly use an external signature file for module extraction in this study, we have extended the tool and integrated it with the Protégé DL Query plugin [20], by which a signature can be defined through a semantic query which invokes a DL (description logic)-based reasoner. We consider that this provides a powerful and easy to use feature to define and extract a domain specific value set.

In conclusion, we performed a case study for ICD-11 anatomy value set extraction from SNOMED CT. We proposed four different clinical context patterns for the purpose of generating clinically meaningful value sets for ICD-11 anatomy. We evaluated the value sets in terms of domain coverage, granularity and clinical usefulness by defining quantitative measures, which provide effective metrics for helping us to select an approach for satisfying the ICD-11 anatomy use case.

References

1. WHO. Revision of International Classification of Diseases (ICD): http://www.who.int/classifications/icd/ICDRevision/en/index.html.
2. WHO. ICD-11 Alpha – Content Model Reference Guide: http://icat.stanford.edu/.
3. Pathak J, Jiang G, Dwarkanath SO, Buntrock JD, Chute CG. LexValueSets: an approach for context-driven value sets extraction. AMIA Annu Symp Proc. 2008 Nov 6:556-60.
4. CTS2 wiki: http://informatics.mayo. edu/cts2.
5. The LexEVS 6.0 URL: https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/LexEVS_Version_6.0.
6. The OWL 3 API URL: http://sourceforge.net/projects/owlapi/.
7. Sattler U, Schneider T, Zakharyaschev. Which kind of module should I extract? In: DL Home 22nd International Workshop on Description Logics, July 27-30, 2009, Oxford, UK.
8. Chute CG. Distributed biomedical terminology development: from experiments to open process. Yearb Med Inform. 2010:58-63.
9. The IHTSDO URL: http://www.ihtsdo.org/snomed-ct/. Last visited at February 15, 2011.
10. Agreement between IHTSDO and WHO: http://www.who.int/classifications/AnnouncementLetter.pdf.
11. Hahn U, Schulz S, Romacker M; Partonomic reasoning as taxonomic reasoning in medicine. Proc 16th National Conf Artificial Intelligence & 11th Innovative Applications of Artificial Intelligence (AAAI-99/IAAI-99):; 271-276.
12. Schulz S, Hahn U, Romacker M; Modeling anatomical spatial relations with description logics. 2000; AMIA Fall Symposium, 799-783.
13. The IHTSDO. SNOMED CT Clinical Terms User Guide. July 2010 International Release.
14. SNOMED CT anatomy data file: https://csfe.aceworkspace.net/sf/go/doc3132?nav=1.
15. The CORE Problem List Subset of SNOMED CT Download: http://www.nlm.nih.gov/research/umls/Snomed/core_subset.html.
16. Manchester SNOMED CT Module Extraction Tool: http://owl.cs.manchester.ac.uk/snomed/..
17. Rector AL, Brandt S, Schneider T. Getting the foot out of the pelvis: modeling problems affecting use of SNOMED CT Hierarchies in practical applications. JAMIA. 2011. (in press).
18. The Protege4.1. Plugin for SNOMED CT Module Extraction URL: https://sites.google.com/site/ontologymodularit/.
19. The International Classification of Diseases for Oncology (ICD-O) URL: http://www.who.int/classifications/icd/adaptations/oncology/en/.
20. The Protégé DL Query Plugin. http://protégéwiki.stanford.edu/wiki/DLQueryTab.