Exploratory Reverse Mapping of ICD-10-CA to SNOMED CT

Dennis Lee, M.Sc., Francis Lau, Ph.D.
School of Health Information Science, University of Victoria, Victoria, B.C., Canada
dlkh@uvic.ca, fylau@uvic.ca

ABSTRACT

This paper describes the findings of an exploratory study on reverse mapping of ICD-10-CA, the Canadian Adaptation, to SNOMED CT. For this study a set of 5,000 most frequent ICD-10-CA codes from the health ministry of a Canadian province was used. The methods included applying six mapping algorithms to each ICD-10-CA description to find the matching SNOMED CT concepts, and comparing the output against the UK SCT-ICD10 cross map for accuracy. Overall, we found successful SNOMED CT matches for ~63% of the ICD-10-CA codes. Issues requiring further attention include ways to increase successful matches and independent validation of mapping output. This study provides a glimpse of the methods that could lead to a SNOMED CT to ICD-10-CA cross map. It should be of interest to those responsible for secondary use of discharge abstracts in epidemiological and statistical reporting.

INTRODUCTION

The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) is a terminology system used to capture information relating to a patient's condition and care in a consistent manner. Currently, there are ~376000 concepts in SNOMED CT, organized into 19 hierarchies such as clinical finding, observations, body structure and social context. There are another ~1 million commonly used terms to describe these concepts, and ~1.4 million semantic relationships to define the logical connections between concepts [1].

While SNOMED CT is the terminology of choice for capturing details of a clinical encounter, it is considered too fine grained for non-clinical purposes such as the reporting of resource use and billing. Many have advocated the need to link SNOMED CT to established classification systems, such as the International Statistical Classification of Diseases and Related Health Problems Version 10 (ICD-10), that are already used extensively in statistical reporting [2,3]. Currently there is a cross map from SNOMED CT to ICD-10 in the UK, and one to ICD-9-CM (Clinical Modification) in the United States. Neither of these maps have been validated externally, and no map exists for ICD-10-CA, the Canadian Adaptation. There are other cross maps that have been created for specific domains including the SNOMED-to-ICD-O map for oncology, the SNOMED-to-LOINC map for laboratory test results, and those for nursing terminologies. Otherwise there is limited experience in cross mapping from SNOMED CT to existing classification systems to facilitate secondary uses.

In this paper, we describe the initial findings of an exploratory study to create a reverse map from ICD-10-CA to SNOMED CT. It originated as part of a Master of Science project by the lead author. We contend that reverse mapping could be one way to produce the SNOMED CT to ICD-10-CA cross map. This paper describes the mapping algorithms and process used, the key results on matches found, and the lessons and implications from the study.

METHODS

Overview of ICD-10-CA

The ICD-10-CA is an enhanced version of the ICD-10 published by the World Health Organization (WHO). The ICD-10-CA has 23 chapters and is used for classifying morbidity, diseases, injuries and causes of death in Canada. It also covers non-disease situations and conditions that pose a risk to health including occupational and environmental factors, lifestyle and psycho-social circumstances. The ICD-10-CA has an alphanumeric coding format of 3-6 characters. The major difference between ICD-10 and ICD-10-CA is that the latter has two additional chapters: XXII on morphology of neoplasms and XXIII on provisional codes for research and temporary assignment. There are also minor changes in some chapters in the form of addition, subdivision, deletion and revision of selected ICD codes [4].

Source Mapping Terms

For this study, we obtained a set of 5,000 most frequently reported ICD-10-CA codes and their long descriptions for the fiscal year of 2005/06 from the health ministry of a Canadian province. These source mapping terms were from inpatient separations in acute care settings including designated sub-acute care facilities for patients that require more care and time before returning home. The profile of the discharge abstracts for the 5,000 ICD-10-CA codes selected for the study is in Table 1.

Description	**Count**
Total separations 2005/06 in province	364,977
Total diagnosis codes reported	1,481,285
Average no. of codes reported per separation	4.1
Total discrete diagnosis codes (all)	10,529
Frequency of top 5,000 diagnosis codes	1,460,730
% of total diagnosis in top 5000 codes	98.6%
% of total discrete diagnosis in top 5000 codes	47.5%
Total discrete most responsible diagnosis codes	6,651

Table 1. Profile of the Discharge Abstracts

Mapping Algorithms

After conducting a detailed review of the literature on cross mapping of terminology systems, we adopted five related mapping algorithms and created Web-based versions of these algorithms in to find matching SNOMED concepts for each of the ICD-10-CA descriptions in the data set [5]. Four of the algorithms are lexical techniques for exact-match, match-all-words-only, match-all-words and partial-match. The fifth is semantic matching that involves retrieving the current concepts based on entries in the SNOMED historical relationship table if the initial concepts found are inactive. These mapping algorithms are summarized in Table 2.

Algorithm	**Explanation**
1. Exact match	Exact string match where all words are same and in same sequence for both source and target terms, including punctuation
2. Match all only	String match where all words are same but not necessary in same order; additional words not allowed in target term
3. Match all	String match where all words are same but not necessary in same order; additional words allowed in target term
4. Partial match	String match where one or more words in source term is found in target term
5. Semantic match	For inactive concepts found use historical relationships of Was-A Same-As, May-Be-A, Replaced-By to find current concepts
6. Unmappable	Assigned when no match is found

Table 2. Mapping algorithms used in this study

Normalization Steps

In addition to using the original SNOMED CT terms and the ICD-10-CA long descriptions in mapping, we normalized all of these original terms to remove "noise" such as genitives and spelling errors using the Unified Medical Language System (UMLS) normalization steps, as shown in Table 3a [6]. To improve successful mapping, we expanded step-2 to remove both "stop words" and "exclude words," as well as SNOMED prefixes, shown in Table 3b. For step-5 we included both the lookup and stemming methods to uninflect the phrase. The lookup method uses the UMLS SPECIALIST Lexicon's inflection table with ~1 million entries, whereas the stemming method uses the computational technique first published by Porter Stemming that reduces word variants to a single canonical form [7,8].

Steps 1 to 6	**Example**
Remove genitive	Hodgkin***'s*** disease, NOS → Hodgkin diseases, NOS
Remove stop words	Hodgkin diseases, ***NOS*** → Hodgkin diseases,
Convert to lowercase	***H***odgkin diseases, → hodgkin diseases,
Strip punctuation	hodgkin diseases***,*** → hodgkin diseases
Uninflect phrase	hodgkin disease***s*** → hodgkin disease
Sort words	***hodgkin*** ***disease*** → disease hodgkin

Table 3a. UMLS six normalization steps[7, slide 20]

Step-2	**Explanation**
Stop words	Frequent short words that do not affect the phrase: and, by, for, in, of, on, the, to, with, no, and (nos)
Exclude words	Words that may change meaning of the word but if ignored help to locate a term otherwise missed: about, alongside, an, anything, around, as, at, because, before, being, both, cannot, chronically, consists, covered, does, during, every, find, from, instead, into, more, must, no, not, only, or, properly, side, sided, some, something, specific, than, that, things, this, throughout, up, using, usually, when, while, without
SNOMED Prefixes	[X] – concepts with ICD-10 codes not in ICD-9 [D] – concepts in ICD-9 XVI and ICD-10 SVII [M] – morphology of neoplasm concepts in ICD-O [SO] – concepts in OPCS-4 chapter Z in CTV3 [Q] – temporary qualifying terms from CTV3 [V] – concepts in ICD-9 and ICD-10 on factors influencing health status and contact with health services (V-codes and Z-codes)

Table 3b. Expanded UMLS normalization step-2

Reverse Mapping Process

The reverse mapping of ICD-10-CA terms to SNOMED CT concepts involved cycling through the mapping algorithms one at a time to find the best candidate SNOMED CT concepts as the target terms. For each algorithm we always started with the original terms, then the UMLS normalized terms, followed by the stemmed terms. In each cycle, we would review the candidate concepts found to see if it was a match, and if so, what type of match it was based on the algorithm applied. When no matching concepts were found, we would label the term as unmappable. Our experience with the matching techniques was that, the sooner we could find a match in the cycle, i.e. first-match, the greater confidence we would have that the candidate concept is appropriate. The preferred order of matched terms was always exact-match first, match-all-only, then match-all, with partial-match last. Whenever inactive concepts were found a semantic-match was done to find the current concepts through their historical relationships. During mapping we tallied frequency statistics on the different types of matches with summary/detailed outputs. Only the first-matches were counted to determine the effectiveness of each mapping algorithm.

Comparison with UK SCT-ICD10 Map

To determine the accuracy of the mapping results from this study, we compared our output with the UK SNOMED CT to ICD-10 (SCT-ICD10) cross map. To do so, the 5,000 ICD-10-CA codes were matched with the *TargetCodes* of the *SCT_CrossMapTargets* table from the July 2007 version of the IHTSDO distribution set [1]. While the UK cross map is from SNOMED CT to ICD-10 and not ICD-10-CA, the two ICD versions share many similar codes. Thus, if the ICD-10-CA code was found among the *TargetCodes* of the UK map, we would look up the *SCT_CrossMaps* table to find the corresponding SNOMED concepts. If multiple similar SNOMED concepts were found, they would be filtered to include only the unique SNOMED concepts. Each of the concepts found were then compared with our mapping output from matches found by the exact-match, match-all-only and match-all algorithms.

RESULTS

Summary of Mapping Output

Of the 5,000 ICD-10-CA descriptions used in this study, we were able to match 1,619 source ICD terms (32.38%) to 2,625 target SNOMED concepts by the exact-match technique. Next, we matched 63 ICD terms (1.26%) to 87 SNOMED concepts by match-all-only; another 1,478 ICD terms to 4,829 concepts by match-all; and 1,839 ICD terms to ~25 million concepts by partial-match. One ICD term *C8800 Waldenstr* was umappable. A summary of the mapping output by match-type is shown in Table 4.

Match Type	Source	Target	Percentage
Exact match	1,619	2,625	32.38%
Match all only	63	87	1.26%
Match all	1,478	4,829	29.56%
Partial match	1,839	24,950,238	36.78%
Unmappable	1	0	0.02%
Total	**5,000**	**24,957,779**	**100.00%**

Table 4. Summary of Mapping Output

Detailed Analysis of Mapping Output

Each ICD term was cycled through all the matching techniques to determine the number of candidate target SNOMED concepts found for each match type. The first-match reported for each match type excluded the target concepts already identified in previous iterations to avoid duplicate counting. We tracked not only the total matches but also which technique found the first match. The output produced suggested exact-match, match-all-only and match-all could be considered as successful matches, since they returned one or more identical or similar SNOMED concepts based on the ICD term provided. The number of first-matches found for these match types by ICD Chapter are shown in the Appendix. One can see that the percentages of matches were very low for Chapters *IV Endocrine, nutritional and metabolic diseases* at 36%; *XIII Diseases of the musculoskeletal system and connective tissue* at ~36%; and *XV Pregnancy, childbirth and the puerperium* at ~4%. Of the overall 3,160 ICD terms or ~63% that were mapped to one or more SNOMED concepts, most were found by exact-match and match-all during the first-match. The profiles of first-matches found by each match type are briefly described below.

Exact Match – Table 5 shows 1,237 original ICD terms had exact-matches with 2,064 candidate concepts. Another 364 ICD terms had exact-matches with 527 concepts using the UMLS normalized version, and 18 ICD with 34 concepts using the stemmed version. In all, 2,625 candidate SNOMED concepts were found, which means that there were multiple exact matches for some of the ICD terms.

Exact Match	First Match	Target
Original Term	1,237	2,064
UMLS Version	364	527
Stemmed Version	18	34
Total	**1,619**	**2,625**

Table 5. Exact match output

Match All Only – Table 6 shows 33 original ICD terms had match-all-only with 48 candidate concepts; 29 UMLS normalized terms had 37 concepts, and 1 stemmed term had 2 only. In all, 87 candidate SNOMED concepts were found, which means that there were multiple match-all-only for some terms.

Match All Words Only	First Match	Target
Original Term	33	48
UMLS Version	29	37
Stemmed Version	1	2
Total	**63**	**87**

Table 6. Match all only output

Match All Words – Table 7 shows 1,343 original ICD terms had match-all with 4,558 candidate concepts; 114 UMLS normalized terms had 217 concepts, and 21 stemmed terms had 54. In all, 4,829 SNOMED concepts were found, which means that there were multiple match-all for some terms.

Match All Words	First Match	Target
Original Term	1,343	4,558
UMLS Version	114	217
Stemmed Version	21	54
Total	**1,478**	**4,829**

Table 7. Match all words output

Partial Match – Table 8 shows 1,839 ICD terms had partial-matches with 25 million SNOMED concepts. We found the results of partial matches to be more unpredictable than the previous match types. If a source term was long and contains common words such as *disorder* or *procedure*, the results returned could be numerous as only one word from the source term needed to be present in the target term.

Partial Match	First Match	Target
Original Term	1,839	24,950,238
UMLS Version	0	0
Stemmed Version	0	0
Total	**1,839**	**24,950,238**

Table 8. Partial match output

Comparison with SCT-ICD10 Map

Six comparisons were made between our mapping output and the UK map to see if: (a) both contained the same results; (b) both contained similar results; (c) both contained dissimilar results; (d) only UK map contained the results; (e) only our mapping output contained the results; (f) both had unmappable results. The overall results are shown in Table 9. Only (b), (c) and (f) are illustrated in this paper.

Type of comparison	Frequency	Percentage
Contained exactly same results	11	0.22%
Contained similar results	2,401	48.02%
Contained dissimilar results	122	2.44%
UK map with results only	896	17.92%
Mapping outputs with results only	370	7.40%
Both had unmappable results	1,200	24.00%
Total	**5,000**	**100.00%**

Table 9. Comparing UK map and mapping outputs

Similar Results - Where both maps contained similar results, the UK map usually had more mapped terms than our output, as shown in Table 10. An example is with the ICD term *Q61.2 Polycystic kidney, autosomal dominant* where the UK map had six SNOMED concepts but only four in ours.

Description			Total
UK map had more results than mapping outputs			2,125
Mapping outputs had more results than UK map			224
UK and mapping outputs had same no. of results			63
Total			**2,401**
ConceptId	**Fully Specified Name**	**UK**	**CA**
66091009	Congenital disease (disorder)	√	
204955006	Polycystic kidney disease	√	
204962002	Multicystic kidney (disorder)	√	
28728008	Polycystic kidney disease, adult type (disorder)	√	√
253878003	Adult type polycystic kidney disease type I (disorder)	√	√
253879006	Adult type polycystic kidney disease type II (disorder)	√	√
274567009	[EDTA] Polycystic kidneys, adult type (dominant) associated with renal failure (disorder)		√

Table 10. Comparing both with similar results

Dissimilar Results – Where both had dissimilar results, our output were more specific as each concept must contain all the words in the source term. For 100 (82%) of these terms the UK map had more candidate concepts; for 9 terms (7.4%) both had same number of concepts; whereas for 13 (10.7%) our mapping output had more concepts. An example is the ICD term *S597 Multiple injuries of forearm*, shown in Table 11, where both maps had four concepts but none are similar.

ConceptId	Fully Specified Name	UK	CA
122549002	Injury (disorder)	√	
125596004	Injury of elbow (disorder)	√	
210557006	Severe multi tissue damage lower arm (disorder)	√	
210558001	Massive multi tissue damage lower arm (disorder)	√	
210860005	Injury of multiple blood vessels at forearm level (disorder)		√
211290004	Multiple superficial injuries of forearm (disorder)		√
212308001	Injury of multiple nerves at forearm level (disorder)		√
212464002	Injury of multiple muscles and tendons at forearm level (disorder)		√

Table 11. Comparing both with dissimilar results

Unmappable Results – These were in almost every ICD chapter but most notable in *XVII: Congenital malformations, deformations and chromosomal abnormalities; XIX: Injury, poisoning and certain other consequences of external causes; and XIII: Diseases of the musculoskeletal system and connective issue* (Table 12). It is possible these ICD terms have further refinement making it difficult to find concept and lexical matches. An example is the ICD-10-CA term *O2450 Pre-existing Type 1 diabetes mellitus arising in pregnancy,* which could be refined as: *delivered with or without antepartum condition (1), delivered with postpartum complication (2), or antepartum condition or complication (3).*

Chapter	Range	Freq	%
XVII: Congenital malformations, deformations, and chromosomal abnormalities	Q00-Q99	292	24.33%
XIX: Injury, poisoning and certain other consequences of external causes	S00-T98	278	23.17%
XIII: Disease of the musculoskeletal system and connective tissue	M00-M99	207	17.25%
IV: Endocrine, nutritional and metabolic diseases	E00-E90	119	9.92%
XX: External causes of morbidity and mortality	V01-Y98	60	5.00%
		956	79.67%

Table 12. Unmappable ICD-10-CA terms

DISCUSSION

Lessons and Issues

This study was our initial effort to apply a set of mapping algorithms on a set of ICD-10-CA terms to find the matching target SNOMED concepts. Our output showed most of the matches were found using the exact-match and match-all algorithms. The match-all-words-only algorithm did not add a great deal to the number of matches found, and the partial-match was considered too unpredictable with respect to the candidate target concepts returned. Due to space limitation, we did not report on additional matches found after normalization with UMLS and stemming techniques were applied to the original ICD terms, or those found by semantic matching.

A major issue is how one should define "successful match." In our output we had just over 60% of the matches found by exact-match and match-all, which we reviewed and deemed correct. However, more formal validation preferably by an independent source is needed. While our results showed successful matches in only ~63% of the 5,000 ICD-10-CA codes, we were surprised to find the UK cross map had similar successful matches of ~68% against the same 5,000 ICD-10-CA codes (see Table 9). Equally intriguing were the different matches found between the two maps. Almost 50% of the concepts found were similar but not identical, whereas ~20% were dissimilar or found only in the UK map. One possible explanation is the minor differences that exist between ICD-10 and ICD-10-CA with respect to the addition, subdivision, deletion and revision made in some ICD-10-CA chapters. Another is that a concept-based method was used to create the UK cross map, which seemed to outperform the lexical techniques in this study. One possible solution to improve mapping precision is to combine methods, such as the use of semantic and lexical mapping between SNOMED CT and ICD-9-CM by Fung.[9]

Another issue is the extent that our semi-automated matching algorithms can aide in the cross-mapping process by health records staff when encoding the inpatient discharge abstracts. The current abstracting process is mostly an intellectual and manual exercise. As such, explicit cross-mapping guidelines need to be established, including the use of any computer-based mapping tools, to improve this abstracting process. With our mapping algorithms, a consensus-based process is needed for the health record staff to verify the accuracy of the ~63% successful matches. Guidelines are also needed to reconcile the remaining ~37% partially-matched terms.[2,10]

Still, we contend there is merit in exploring the use of reverse mapping with lexical algorithms to identify candidate SNOMED concepts for a given set of ICD-10-CA terms. Our next steps are to enhance the mapping algorithms to include contexts, incorporate these algorithms into the abstracting process, and conduct further field evaluation. Last, the idea of applying reverse mapping to identify candidate SNOMED CT concepts for a set of mapping terms can be a helpful approach when creating a cross map from SNOMED CT to another terminology system.

Implications

This study provides a glimpse of the feasible mapping methods that could eventually lead to a SNOMED CT to ICD-10-CA cross map for Canada. We believe the intent, methods and results of this current study should be of interest to those responsible for secondary use of patient discharge abstracts in epidemiological and statistical reporting. The notion of reverse mapping is also highly generalizable to include the encoding of local terms that already exist in legacy systems within many health organizations to a reference terminology such as SNOMED CT.

Acknowledgments

We wish to thank the Provincial Ministry that provided the 5,000 ICD-10-CA codes for the study. We also thank Ms. Robyn Kuropatwa in facilitating the process to obtain the ICD codes from the ministry. Funding support for this work was provided by the Canadian Institutes for Health Research through its Strategic Training Initiative. Note that the views presented in this paper are those of the authors only and do not represent the official position of any Canadian government agencies.

REFERENCES

1. IHTSDO, International Health Terminology Standards Development Organization. *SNOMED Clinical Terms Technical Reference Guide*. International Release, July 2007.
2. Bowman S. Coordination of SNOMED CT and ICD-10: Getting the Most out of Electronic Health Record Systems. *Perspectives in Health Information Management*, Spring 2005.
3. McBride S, Gilder R, et al. Data mapping. *Journal of American Health Information Management Association* 2006; 77(1): 44-48.
4. CIHI, Canadian Institute for Health Information. Canadian Coding *Standards for ICD-10-CA and CCI for 2006.* Ottawa, Canada. 2006.

5. Lee DHK. *Reverse Mapping ICD-10-CA to SNOMED CT*. UVic Master of Science research project report, Oct 2007. Unpublished.
6. National Library of Medicine. *The SPECIALIST Lexicon*. http://lexsr3.nlm.nih.gov/LexSysGroup/Projects/Summary/lexicon.html
7. Kleinsorge R, Willis J, et al. UMLS Overview – Tutorial T12. *AMIA Annual Symposium* 2006. http://165.112.6.70/research/umls/pdf/AMIA_T12_2006_UMLS.pdf. Jan15/2006.
8. Goldsmith JA, Higgins D, Soglasnova S. *Automatic Language-specific Stemming in Information Retrieval.* Springer-Verlag Berlin Heidelberg 2001.
9. Fung KW, Bodenreider O, Aronson AR, Hole WT, Srinivasan S. Combining lexical and semantic methods of inter-terminology mapping using the UMLS. In Kuhn K. et al. (Eds) *MedInfo 2007*, p605-610. IOS Press, 2007.
10. Vikstrom A, Skaner Y, et al. Mapping of the categories of the Swedish primary health care version of ICD-10 to SNOMED CT concepts: Rule development and intercoder reliability in a mapping trial. *BMC Medical Informatics and Decision Making* 2007;7:9.

Appendix. Mapping Output for top 5,000 ICD-10-CA codes by ICD Chapter

Chapter	Title	Range	Source	Exact	Only	All	Total	Percent
I	Certain infections and parasitic disease	A00-B99	136	47	2	57	106	77.94%
II	Neoplasms	C00-D48	343	174		58	232	67.64%
III	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	D50-D89	80	35	1	20	56	70.00%
IV	Endocrine, nutritional and metabolic diseases	E00-E90	225	56	1	24	81	36.00%
V	Mental and behavioural disorders	F00-F99	218	66	3	141	210	96.33%
VI	Diseases of the nervous system	G00-G99	196	75	1	56	132	67.35%
VII	Diseases of the eye and adnexa	H00-H59	89	56	3	18	77	86.52%
VIII	Diseases of the ear and mastoid process	H60-H95	42	24		11	35	83.33%
IX	Diseases of the circulatory system	I00-I99	279	136	1	74	211	75.63%
X	Diseases of the respiratory system	J00-J99	165	67	4	41	112	67.88%
XI	Diseases of the digestive system	K00-K93	276	136	9	56	201	72.83%
XII	Diseases of the skin and subcutaneous tissue	L00-L99	105	42		20	62	59.05%
XIII	Diseases of the musculoskeletal system and connective tissue	M00-M99	383	78	1	61	140	36.55%
XIV	Diseases of the genitourinary system	N00-N99	226	120	3	48	171	75.66%
XV	Pregnancy, childbirth and the puerperium	O00-O99	313	5	1	6	12	3.83%
XVI	Certain conditions originating in the perinatal period	P00-P99	169	57	17	47	121	71.60%
XVII	Congenital malformations, deformations, chromosomal abnormalities	Q00-Q99	205	105	2	57	164	80.00%
XVIII	Symptoms, signs and abnormal clinical and laboratory findings not elsewhere classified	R00-R99	181	99	2	52	153	84.53%
XIX	Injury, poisoning and certain other consequences of external causes	S00-T98	691	175	8	169	352	50.94%
XX	External causes of morbidity and mortality	V01-Y98	297	9	4	249	262	88.22%
XXI	Factors influencing health status and contact with health services	Z00-Z99	333	29		199	228	68.47%
XXII	Morphology of neoplasms	8000/0-9989/1	28	28			28	100.00%
XXIII	Provisional codes for research and temporary assignment	U00-U99*	20			14	14	70.00%
	Total		**5,000**	**1,619**	**63**	**1,478**	**3,160**	**63.20%**