

# Linguistic Mapping of Terminologies to SNOMED CT

Yefeng Wang<sup>1</sup>, Jon Patrick<sup>1</sup>, Graeme Miller<sup>2</sup>, Julie O'Halloran<sup>2</sup>

<sup>1</sup>School of Information Technology, University of Sydney, Sydney, Australia

<sup>2</sup>Family Medicine Research Centre, University of Sydney, Sydney, Australia

## ABSTRACT

*In this study we develop some linguistic bases for mapping between terminologies and demonstrate their application on mapping ICPC-2 PLUS to SNOMED CT (SCT). The Unified Medical Language System (UMLS) metathesaurus mapping, which utilises the links between ICPC-2 PLUS and SCT terms in the UMLS library mapped 46.5% of ICPC-2 PLUS terms to SCT. Lexical mapping explored the lexical similarities between terms in these two terminologies, and mapped 60.3% of ICPC-2 PLUS terms overall. Post-coordination of remaining unmapped terms was performed, allowing one ICPC-2 PLUS term to be mapped into composition with two SCT terms, which gives an increase of about 20% in mapped terms. Overall we have mapped 80.58% of ICPC-2 PLUS terms to SCT terminology with different level of accuracy. A manual review of the mapping shows that about 90% of string-based mappings are accurate. Unmapped terms and mismatched terms are due to the differences in the structures between ICPC-2 PLUS and SCT. Terms contained in ICPC-2 PLUS but not in SCT caused a large proportion of failures in the mappings.*

## INTRODUCTION

Effective information retrieval across information systems in health are limited by the lack of semantic interoperability between terminologies used by sectors in the health system. The use of multiple terminologies and ad hoc modifications to standard schemes prevent users from cross searching multiple repositories, cross-sectoral resources and inter-disciplinary material.<sup>1</sup> In order to co-ordinate this, improved matching between terms is required. The process of terminology mapping refers to an identification of identical concepts or relationships between different terminologies. Terminology mapping is an important step to achieving knowledge sharing. There has been a large amount of effort spent on computerised terminology mapping.<sup>2,5</sup> However the nature of this task makes it is very difficult to automate, because heterogeneous terminologies may reflect fundamentally subtly different conceptualisations of domains by the creators of these terminologies.<sup>6</sup> In this research we aim to develop an algorithm to map ICPC-2 PLUS terms into SNOMED CT (SCT) concepts. This mapping process is semi-automatic,

because it requires humans to verify the results at the end of mapping, but it transforms the time consuming searching and mapping task into an easier selection and validation task.

## BACKGROUND

ICPC-2 PLUS<sup>7</sup> is an interface terminology classified to the International Classification of Primary Care Version 2 (ICPC-2). It is developed and maintained by The Family Medicine Research Centre (FMRC) of the University of Sydney. The ICPC-2 is a classification designed for general practice and primary care divided into 17 sections such as Musculoskeletal, Neurological, Eye and Blood etc. The ICPC-2 PLUS is an extension to the ICPC-2 and a version used in Australia. It provides a useable coding system for symptoms, diagnoses (problem labels), past health problems and the processes of care for use in age-sex disease registers, morbidity registers and full electronic health records in primary care. It currently contains over 7,000 terms that are commonly used in Australian general practice.

SCT is the most comprehensive biomedical terminology in the world. It contains more than 360,000 concepts which are connected by complex relationship networks. It has been used in coding clinical records and the Australian government is proposing to adopt it for describing certain aspects of clinical encounters and coding clinical records. This decision creates a need to map ICPC-2 PLUS codes to SCT codes. To complete this task in a reasonable amount of time and improve accuracy, some computational methods of matching concepts are needed to assist humans to do the job.

The extensive research done in terminology mapping has had the goal of developing effective automated methodologies for mapping. The main approaches include lexical matching, concept matching and structural matching. A number of linguists have attempted to make use of linguistic information such as lexical similarity and semantic similarities.<sup>8</sup> Other approaches have been developed recently using structural information to map between terminologies. Mork and Bernstein<sup>9</sup> modified a genetic terminology mapping algorithm for mapping human anatomy, using lexical similarities and structure

similarities. However, medical terminologies are different from general terminologies, and they are organised in different axis. The structural mapping is only moderately effective. Moreover, the medical terminology contains hundreds of thousands of concepts, so searching through all concepts is inefficient. The Unified Medical Language System (UMLS) has been widely used as a knowledge resource in many medical terminology mapping works. Fung and Bodenreider<sup>10</sup> derived an algorithm to find candidate mappings between any two terminologies inside the UMLS making use of synonymy, explicit mapping relations and hierarchical relationships. Post-coordination can be used to map terms to compositions of two or more concepts. Elkin and Brown<sup>11</sup> developed a technique for discovering and formalising the implicit semantic relationships between SNOMED Reference-Terminology (SNOMED-RT) and International Classification of Disease Version 9 Clinical Modification (ICD9-CM). Julie Green and colleagues evaluated an existing model for structured recording of heart murmur findings<sup>12</sup>. They use the Interprets and Has interpretation concepts in SCT with a grouping mechanism for roles to represent murmur characteristics and attribute values.

## METHODS

### Mapping using UMLS

The Unified Medical Language System (UMLS)<sup>13</sup> is a knowledge source that provides the mapping between different terminologies. This is done by incorporating different medical terminologies into a Metathesaurus organized on the basis of a “concept”. There are over one million concepts, 2.8 million distinct strings from over 100 source terminologies. The 2005AB version of the UMLS contains ICPC-2 PLUS 2000 Version and SCT terminology, which are the terminologies we need to map.

The most direct mapping method used was to utilise the link provided by the UMLS (Unified Medical Language System) Metathesaurus<sup>2</sup>. The UMLS is organised by concepts, and one of its primary purposes is to connect different names for the same concept from many different vocabularies. Similar terms in different vocabularies are implicitly connected by a unique concept identifier. The idea of our approach is to find the terms in these two terminologies that share a common concept unique identifier (CUI) in UMLS. Every term in the UMLS is represented in a concept structure. The concept structure contains concept identifiers, concept names, their language, and vocabulary source. This information is organised in the Concept Names and Sources File. We make use of the common CUI in this file to map terms.

The UMLS mapping requires the latest UMLS Metathesaurus version to get the best performance, since the content of the UMLS, and its source vocabularies, are refined and updated regularly. Current experiments were conducted on the 2005AB version which contained a

version of ICPC-2 PLUS from 2000, and a version of SCT from 2002. The version of ICPC-2 PLUS in UMLS accounts for only 87% of terms currently available in the terminology. ICPC-2 PLUS has since been updated in the UMLS to the most current version, and we therefore expect a larger number of mappings will be discovered when we use the latest version of UMLS.

### String-Based Mapping

The theory behind string-based matching is that most terminologies have lexical similarity in their vocabularies, for describing the same concepts, as the natural languages underlying the vocabularies are the same. Four string based mapping techniques are used.

#### *Normalised Term Matching*

Before comparing the string, the terms from both terminologies are normalised using natural language processing techniques. Firstly, words within parentheses are removed. This removed the suffix attributes in SCT concepts. Then the terms are tokenised into atomic forms and converted into lowercase. Stop words such as “a”, “the”, “of”, “NOS” etc. and punctuation are removed from multi-word expressions. A morphological process is performed on the remaining terms to remove the inflections. Then some common lexical variations of the terms are generated using the Specialist Lexicon<sup>13</sup> in UMLS. Finally the remaining words are sorted in alphabetical order. Then the normalised terms are matched using exact string matching method. An example shows the SCT concept “235856003 *Disease of liver (disorder)*” is normalized to “*disease liver*” and can be mapped to ICPC-2 PLUS term “D97002 *Disease;liver*”.

#### *Expanded Term Matching*

There are two kinds of abbreviations found in ICPC 2-PLUS terms. One is acronyms such as “IUCD” which stands for “*Intra-Uterine Contraceptive Device*” and the other is abbreviations due to space limitations e.g. “*musculo*” for “*musculoskeletal*”. These abbreviations cause mismatches in the string matching process, therefore abbreviations are expanded to their full forms. In the first case, a list of acronym to full form mapping is created using the abbreviation list in ICPC 2-PLUS user’s guide. In the second case, we adapt the information in the natural language description of the term held in ICPC 2-PLUS to expand the abbreviations. The full form terms are then mapped using the string matching method.

#### *Substring Term Matching*

To increase the matching coverage, substring matching is also performed. The pair of the terms are matched/mapped if the normalized and expanded ICPC2 term is a substring of the SCT term. This allows a specific term to map to a general term, for example, the term *chronic pain* is a substring of *chronic back pain*.

Matching Method	ICPC-2 PLUS Term	SNOMED CT Term
Normalised String Matching	L81030 Haemarthrosis;ankle	202415003 Hemarthrosis of the ankle (disorder)
Expanded String Matching	P23008 Disorder;opposit	18941000 oppositional defiant disorder (adolescent)
Substring Term Matching	D21005 Feeling (of);choking	62014003 Adverse reaction to drug (disorder)
WordNet Lexicon Matching	D21005 Feeling (of);choking	373909009 Choking sensation (finding)

Table 1. Example of string based mapping

#### WordNet Lexicon Matching

This matching approach uses thesauruses to explore the syntactical variation and semantic meaning of the word constituents. The WordNet synset<sup>14,15</sup> was used to provide semantic and syntactic information about the term. The WordNet synset contains a list of synonymous terms for a word constituent, which allows the mapping of “heart disease” into “cardiac disease” because “heart” and “cardiac” are synonyms. WordNet also provides the derivationally related terms for a given word which can be used for searching. For example, the word “fever” is linked to its related adjectives “feverish” and “feverous”.

Table 1. shows some examples of string-based mappings.

#### Post-coordination Mapping

There are 13,383 records in the Concept Names and post-coordination mapping process aims to map a pre-coordinated ICPC-2 PLUS term to compositions of two or more SCT concepts. This algorithm consists of three steps. Firstly, we break the ICPC-2 PLUS term into atomic terms. This step includes term normalisation, term expansion and breaking the text into separate words. Then we map each atomic term to the SCT atomic concepts. The atomic term mapping is based on the longest string match. Finally, we find the relationship between the SCT concepts by matching the relationship patterns<sup>16</sup> we discovered in SCT. We aim to map two kinds of post-coordination in SCT, the Qualification and Combination. Table 2 shows some examples of post-coordination.

Source Term	Post-coordination
Pain;mouth	22253000 pain (clinical finding) + 21082005 entire mouth region (body structure) : relationship type = 363698007 finding site (attribute)
Referral;radiologist	3457005 patient referral (procedure) + 66862007 radiologist (occupation) : relationship type = 370131001 recipient category (attribute)
Dislocation;knee;simple	13673007 Simple (qualifier value) + 129156001 Traumatic dislocation of knee joint (Clinical Finding) : relationship type = 246100006 onset (attribute)

Table 2. Example of post-coordination mapping

## RESULTS

This section reports the mapping results of each method. A large set of the UMLS mapping results and string-based mapping results have been evaluated by human experts, and only the best candidate mapping is considered as the correct mapping.

#### UMLS Mapping Results

There are 13,383 records in the Concept Names and Sources File containing ICPC-2 PLUS terms. It includes the active ICPC-2 PLUS terms, inactive ICPC-2 PLUS terms, synonyms, duplicates and language variations. By eliminating the synonyms, duplicates, and language variations, 6,502 terms are currently in active status, which is 87.75% of the ICPC-2 PLUS vocabulary. These terms are mapped to 6,141 unique Concept Unique Identifiers (CUI) in UMLS.

The mapping algorithm has mapped total 3,448 ICPC 2-PLUS terms to SCT concepts through 6,557 Common Unique Identifiers in UMLS, which is an average of 1.9 mappings per ICPC-2 PLUS term.

ICPC2P Entries in UMLS	13,383
Active ICPC 2-PLUS terms in UMLS	6,502
Number of mapped CUI	6,141
ICPC 2-PLUS term mapped to SNOMED	3,448
Number of ICPC-SCT Mapping Candidates	6,557
Number of best-fit mapping	3,326
Average number of mapping per term	1.9

Table 3. Results of mapping using UMLS Metathesaurus

The UMLS mapping matches have been assessed by human experts, and the matches are selected on a one to one map of “best-fit”. All context dependent concepts in SCT were excluded as well as legacy concepts. Some matches are of questionable validity which is due to inappropriate ICPC-2 PLUS mapping to UMLS concepts, and some mappings are reasonable lexical or concept matches but they are categorical mismatches. Only the “best-fit” matching candidates are considered as the correct matches, the remainder of the matches are incorrect mappings. Overall, the UMLS mapping algorithm found 6,557

mapping candidates. 3,326 (50.72%) mapping candidates are best-fit mappings, and 96.49% of the mapping candidates have at least one best-fit mapping.

### String-Based Mapping Result

A total of 3,266 ICPC-2 PLUS terms were mapped to SCT terms using normalized string matching. This matching method generated a total 3565 mapping candidates, on average, 1.2 matches per matched terms. The majority of matched terms are single worded terms and multi-word expressions. Some terms with different spelling variations are also mapped. The Expanded String Matching further mapped 304 terms. It effectively increased the number of mappings in chapter M of ICPC-2 PLUS, because most of the terms in this chapter were compressed to short form, however, the average mappings per term increased to 1.33. Synonym Matching is not very effective and only gave a 1% increase in mapping coverage. Most of the Substring Matching results were one to many, and the average number of matches per term increased to 24.88. Overall,

the string-based term mapped 60.34% of ICPC-2 PLUS terms to SCT concepts.

The normalized string matching results were evaluated by an expert. When she questioned the validity of a term, she discussed it with an senior staff member and they came to an agreement. Similar to the UMLS evaluation, only the “best-fit” matches were considered as correct matches. 3,031(92.8%) terms have at least one correct mapping candidate. Among the 3,770 mapping candidates, 3,565(94.25%) were correct mappings.

Several mismatched terms were due to coordination of the terms, the term is connected with conjunctions, slashes etc. such as the term “Splint/immobilise; nerve”. Categorical mismatches occur when the source term and target term have strong lexical similarity but belong to different categories. For example the ICPC 2-PLUS term “A59007: Pain management” is mapped to SCT concept 394882004 pain management (speciality), whereas it should be matched to 278414003 pain management (procedure).

Matching Method	Matched	Candidates	%age	Newly Mapped	Inc. %age
Normalized String Matching	3266	3770	44.08%	-	-
Expanded String Matching	3570	4731	48.18%	304	4.10%
WordNet Matching	3662	5321	49.42%	92	1.24%
Substring Matching	4471	108953	60.34%	809	10.88%

Table 4. String-based mapping results.

### Post-coordination Mapping Result

We excluded the terms that had been mapped in the previous mapping algorithms and performed post-coordination mapping on the remainder of the terms. The remaining set consisted of 3,840 terms. These terms do not have any string matches in SCT terminology or they can't be expressed using one single SCT concept.

Post-coordination Type	# of Mapping	%age
Qualification	343	4.63%
Combination	902	12.17%
Undetermined	255	3.44%
Total	1500	20.24%

Table 5. Post-coordination results

Qualifications are the post-coordinations that have at least one qualifier value concept. The post-coordination type of Combinations occurred when the relationship between the concepts could be identified, but did not include qualifications, while undetermined post-coordinations are those with atomic concepts that had been matched to SCT concepts, but the relationship between the concepts could not be determined. Overall there were 20.24% terms mapped using post-coordination.

## DISCUSSION

As the number of medical terminologies increases, it increases the need for terminology integration. As a result, the demand for rapid and effective computer-assisted terminology mapping has arisen. Computerised mapping systems could significantly reduce human effort, especially when mapping large terminologies. The algorithms used in this study have overall found candidate mappings for 80% of ICPC-2 PLUS terms.

The UMLS has usually been considered as a golden standard. However, it still produces on average 1.9 mappings per term. We think this is due to the synonyms in UMLS and different preferred terms used between countries.

On evaluation, the normalized string matching and expanded string matching were accurate and useful which is about 50% of the ICPC-2 PLUS terms. The substring matching had broader coverage, but resulted in a huge number of mapping candidates. Upon normal inspection, a lot of substring mappings were imprecise. Nevertheless, roughly 10% of the mappings were still accurate. One possibility for reducing the superfluous mapping candidates in string-based mapping was to use the semantic information and categorical information in the SCT hierarchy terminologies to eliminate the irrelevant mappings.

Initially we expected that the structural information of these two terminologies could have provided some useful clues for the matching, however these two terminologies are organised differently. The ICPC-2 PLUS has a biaxial structure and the sections are organised on the body system and social problems, however, SCT is based on 18 key classes. The different organisation of these two terminologies makes it difficult to utilize the structural information.

The use of synonyms in WordNet is not very useful. By looking at the results, we found that the synonymous terms in the SCT descriptions are able to capture most of the synonyms. The results of WN mapping are not as effective as the work done by Mougín<sup>17</sup>, because the matching criteria we used is restrictive and produces less ambiguity in matching candidates.

The results of post-coordination mapping have not yet been evaluated. Nevertheless, the system has demonstrated its ability for automated term decomposition using a combination of string-based mapping techniques. One important phenomenon in post-coordination is the identification of relationships between the mapped terms. This may require description logic generation and more detailed semantic analysis to make sure the matching of two concepts makes sense. We believe that the post-coordination mapping is a way to solve the content completeness problem among different terminologies.

## CONCLUSION

In conclusion, we have mapped about 80.58% of ICPC-2 PLUS terms to SCT concepts with differing levels of accuracy via three automated mapping approaches. This research has demonstrated that automated mapping based on linguistic principles can perform different levels of terminology mapping. The results have shown that some of the mapping methods produce very reliable mapping, while some methods yield boarder coverage, but less convincing selections. The mapping results provide an opportunity to analyse the differences in these two different terminologies. Further refinement of the mapping methods could be done to reduce superfluous and incorrect mapping using structural and categorical information, for example, the elimination of synonym ambiguity. Also, more sophisticated post-coordination mapping could be developed in order to provide more reliable mapping.

## References

1. McCray AT, Loane RF, Browne AC and Bangalore AK, *Terminology issues in user access to Web-based medical information*. Proc AMIA Symp 1999: p. 107-11.
2. Rogers JE, Price C, Rector AL, Solomon WD and Smejko N, *Validating clinical terminology structures: integration and cross-validation of Read The-*

- saurus and GALEN*. In AMIA Fall Symposium, Lake Buena Vista, FL, USA, 1998.
3. Dameron O, Rubín DL and Musen AM, *Challenges in converting frame-based ontology into OWL: the Foundational Model of Anatomy case-study*. in Proc AMIA Annual Symposium 2005: p. 181-5.
4. Fung KW, Hole WT, Nelson SJ, Srinivasan S, Powell T and Roth L, *Integrating SNOMED-CT into the UMLS: An exploration of different views of synonymy and quality of editing*. J Am Med Inform Assoc, 2005. 12(4): p. 486-494.
5. Barrows RC Jr, Cimino JJ and Clayton PD, *Mapping clinically useful terminology to a controlled medical vocabulary*. in Proc Annu Symp Comput Appl Med Care. 1994: p. 211-5.
6. Rector., AL, *Clinical terminology: Why is it so hard?* Methods Inf. Med., 1999. 38(4-5): p. 239-52.
7. ICPC-2 PLUS, An interface terminology classified to the International Classification of Primary Care Version 2. <http://www.fmrc.org.au/icpc2plus/>
8. Noy NF and Musen MA, *PROMPT: Algorithm and tool for automated ontology merging and alignment*. in Proceedings of the National Conference on Artificial Intelligence. 2000: p. 450-5.
9. Mork P and Bernstein PA, *Adapting a Generic Match Algorithm to Align Ontologies of Human Anatomy*. in Proceedings of the 20th International Conference on Data Engineering. 2004.
10. Fung KW and Bodenreider O, *Utilizing the UMLS for Semantic Mapping between Terminologies*. in Proceedings of AMIA Annual Symposium. 2005: p. 266-70.
11. Elkin PL and Brown SH, *Automated enhancement of description logic-defined terminologies to facilitate mapping to ICD9-CM*. Journal of Biomedical Informatics, 2002. 35(5-6): p. 281-8.
12. Green JM, Wilcke JR, Abbott J, and Rees LP, *Development and evaluation of methods for structured recording of heart murmur findings using SNOMED-CT post-coordination*. J Am Med Inform Assoc., 2006. 13(3): p. 321-33.
13. Lindberg DA, Humphreys BL and McCray AT, *The Unified Medical Language System*. Methods of Information in Medicine, 1993. 32: p. 281-91.
14. Fellbaum, C, *WordNet: An Electronic Lexical Database*. The MIT Press. 1998.
15. Burgun, A and Bodenreider O, *Comparing terms, concepts and semantic classes in WordNet and the Unified Medical Language System*. in Proceedings of NAACL'2001 Workshop. 2001.
16. Northfield IL, College of American Pathologists, *Supporting post-coordination. SNOMED CT technical implementation guide July 2003 release*. College of American Pathologists, 2003.
17. Mougín, F., A. Burgun, et al. *Data integration through data elements: Mapping data elements to terminological resources*. Proc Symp on Semantic Mining in Biomedicine, 2006.