e-Health – For Continuity of Care
C. Lovis et al. (Eds.)

doi:10.3233/978-1-61499-432-9-1013

Automatic annotation of ICD-to-MedDRA mappings with SKOS predicates

Gunnar DECLERCK[a,1], Julien SOUVIGNET[a,b], Jean-Marie RODRIGUES[a,b] and Marie-Christine JAULENT[a]
[a]*INSERM, U1142, LIMICS, F-75006, Paris, France; Sorbonne Universités, UPMC Univ Paris 06, UMR_S 1142, LIMICS, F-75006, Paris, France; Université Paris 13, Sorbonne Paris Cité, LIMICS, (UMR_S 1142), F-93430, Villetaneuse, France.*
[b]*Univ. of Saint Etienne, Department of Public Health and Medical Informatics*

Abstract. Robust alignments between ICD and MedDRA are essential to enable the secondary use of clinical data for pharmacovigilance research. UMLS makes available ICD-to-MedDRA mappings, but they are only poorly specified, which introduces difficulties when exploited in an automatic way. SKOS vocabulary can help achieve quality and machine-processable mappings. We have developed an algorithm based on several simple rules which annotates automatically ICD-to-MedDRA mappings with SKOS predicates. The method was tested and evaluated on a sample of ICD-10-to MedDRA mappings extracted from UMLS. The algorithm demonstrated satisfying performances, especially for skos:exactMatch properties, which suggests that automatic methods can be used to improve the quality of terminology mappings.

Keywords. ICD-10, MedDRA, SKOS, UMLS, terminology mapping.

Introduction

Aligning biomedical terminologies is a necessary step to achieve semantic interoperability between health information systems, which determines data integration, data reuse and data sharing [1-2]. Objectives such as secondary use of electronic health record (EHR) data for clinical research, statistical or billing purposes cannot be reached without robust mappings, enabling the – if possible automatic – translation of data from one coding system to another. In the field of drug safety and pharmacovigilance, alignments from the *International Classification of Diseases* (ICD) [3] to the *Medical Dictionary for Drug Regulatory Activities* (MedDRA) are requested. MedDRA is used to record adverse drug reactions (ADR) data and is recommended for the electronic transmission of safety reports [4], while ICD is frequently used to capture clinical information in healthcare settings. An alignment between ICD and MedDRA is for instance a prerequisite to use medical data available in EHR systems for ADR reporting [5] or for mining EHRs or administrative data to detect ADRs [6]. Mappings with ICD-9 and ICD-9-CM were included in previous versions of MedDRA, but they are not maintained and do not cover more recent versions of ICD, such as the widely used ICD-10 and the future ICD-11 (under development).

[1] Corresponding Author. gunnar.declerck@upmc.fr

The most reliable and up-to-date resource currently available to align ICD to MedDRA is undoubtedly the Unified Medical Language System (UMLS) Metathesaurus[2]. Initiatives such as OMOP or BioPortal[3] are also useful, but are less exhaustive. These resources however suffer from important limitations: they do not specify the meaning of the alignment relation being stated, its source and if it has been validated by experts. This is one major inconvenient of UMLS: the codes from several terminologies that are associated with the same UMLS concept (CUI) as possible "synonyms" may correspond to exact match terms, lexical variants, more granular or more general terms, or even to terms related to associated conditions. Such imprecise mappings are only poorly exploitable in an automatic way, leading to unexpected results when used in reasoning algorithms.

Using vocabularies such as SKOS (Simple Knowledge Organization System) [8] appears essential to achieve quality mappings. SKOS provides a precise and machine-processable typology of mapping relations, which enables to exploit the capabilities of semantic reasoning methods when translating data from one terminology to another. For instance, SKOS makes it possible to use logical rules such as subsumption and transitivity to infer non-asserted mappings based on asserted mappings and hierarchical relations between terms inside terminologies. Typically, if A has skos:exactMatch B, and C is-a A, then one can infer that C has skos:broadMatch B. Expressing mappings using SKOS is also essential to evaluate the logical consistency of mappings using semantic reasoning methods, *i.e.* for quality assessment [5].

The problem is that expressing mappings with SKOS requires huge resources if performed manually. To alleviate the manual effort, we have developed an algorithm based on several simple rules, which annotates automatically ICD-to-MedDRA mappings extracted from UMLS with SKOS predicates. In the following, we describe the method we followed to build this algorithm and we present the results in terms of number of SKOS predicates generated and qualitative evaluation. We finally discuss some issues raised by this method and related to the quality of terminologies.

1. Methods

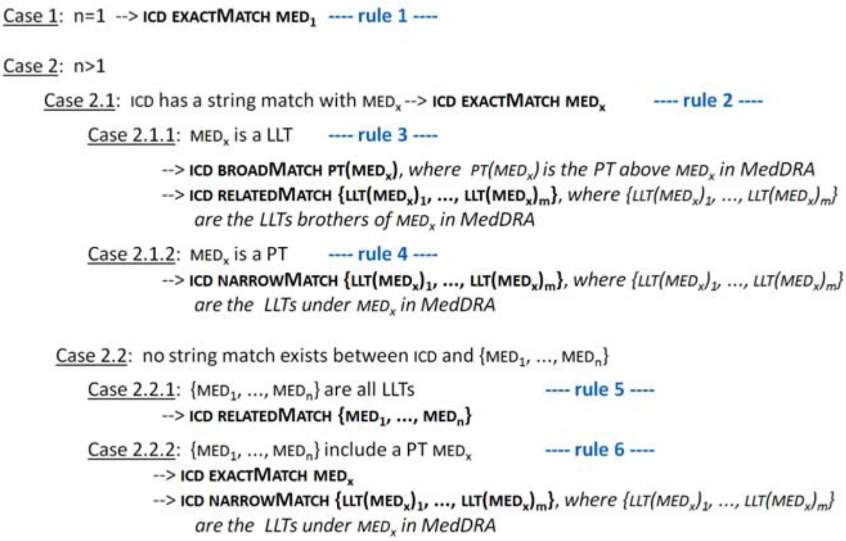
To build the algorithm and evaluate its performances, we used the alignments between ICD-10 and MedDRA 16.0 available in the UMLS 2013AA Release. We extracted from the metathesaurus ICD-10 codes (13,505) and MedDRA codes (93,948); we then performed a junction query associating both codes when the CUI was the same. Because our target use case was to enable automatic EHR data extraction for ADR reporting, we only focused on mappings relating ICD-10 terms to LLT (lowest level term) and PT (preferred term) MedDRA levels, which are recommended for the coding of ADRs [9-10]. ICD-10 codes related to classification titles such as "A00-A09 (Intestinal infectious diseases)" were also ignored (291 codes).

The rules used by the algorithm to generate SKOS predicates are not logical in nature but pragmatic: they have been chosen not because of their formal consistency, but because they appear to work with a satisfying degree of success. They were defined using an iterative method including the following steps: intuitive definition of candidate rules based on a small sample (around fifty) of ICD-to-MedDRA mappings;

[2] http://www.nlm.nih.gov/research/umls/

[3] http://omop.fnih.org/, http://bioportal.bioontology.org/

formalization of these rules in an algorithmic format; testing of the algorithm on a bigger sample; expert evaluation of the results; refinement of the rules. Three iterations were performed to achieve the current version of the algorithm. The algorithm makes use of four out of the five mapping properties available in SKOS: skos:exactMatch, skos:broadMatch, skos:narrowMatch, skos:relatedMatch. skos:closeMatch is not used. The rules used to generate SKOS mapping relations are described below. ICD is an ICD-10 term which is mapped to one or several (*n*) MedDRA terms {MED_1, ..., MED_n}.



Case 1: n=1 --> **ICD EXACTMATCH MED_1** ---- rule 1 ----

Case 2: n>1

Case 2.1: ICD has a string match with MED_x --> **ICD EXACTMATCH MED_x** ---- rule 2 ----

Case 2.1.1: MED_x is a LLT ---- rule 3 ----

--> **ICD BROADMATCH PT(MED_x)**, *where PT(MED_x) is the PT above MED_x in MedDRA*

--> **ICD RELATEDMATCH {LLT(MED_x)$_1$, ..., LLT(MED_x)$_m$}**, *where {LLT(MED_x)$_1$, ..., LLT(MED_x)$_m$} are the LLTs brothers of MED_x in MedDRA*

Case 2.1.2: MED_x is a PT ---- rule 4 ----

--> **ICD NARROWMATCH {LLT(MED_x)$_1$, ..., LLT(MED_x)$_m$}**, *where {LLT(MED_x)$_1$, ..., LLT(MED_x)$_m$} are the LLTs under MED_x in MedDRA*

Case 2.2: no string match exists between ICD and {MED_1, ..., MED_n}

Case 2.2.1: {MED_1, ..., MED_n} are all LLTs ---- rule 5 ----

--> **ICD RELATEDMATCH {MED_1, ..., MED_n}**

Case 2.2.2: {MED_1, ..., MED_n} include a PT MED_x ---- rule 6 ----

--> **ICD EXACTMATCH MED_x**

--> **ICD NARROWMATCH {LLT(MED_x)$_1$, ..., LLT(MED_x)$_m$}**, *where {LLT(MED_x)$_1$, ..., LLT(MED_x)$_m$} are the LLTs under MED_x in MedDRA*

The algorithm works as follows: for a given ICD code, we get a skos:exactMatch if there is only one MedDRA candidate (1), or, in case several MedDRA candidates are available (2), if there is an exact string match (2.1) or a PT (2.2.2). Any remaining MedDRA candidate in a higher level of MedDRA hierarchy is annotated as a skos:broadMatch (2.1.1), in the same level as a skos:relatedMatch (2.1.1 and 2.2.1), in a lower level as a skos:narrowMatch (2.1.2 and 2.2.2).

Some rules (rules 3, 4 and 6) are based on the assumption that the LLTs subsumed by a PT express *more precise* medical concepts, *i.e.* it is assumed that there is always a skos:broadMatch relation between LLTs and PTs, which is not always true since some LLTs are pure synonyms or lexical variants of the PT they relate to [11-12]. This assumption was adapted to the goal of getting a skos:exactMatch mapping for the largest possible number of ICD-10 terms. As a consequence, we also consider that LLTs refer to *different* medical conditions (rules 3 and 5), which, again, is not always true. We will come back to this issue in the Discussion.

2. Results

Using the method described above, 11,647 ICD-to-MedDRA mappings were extracted from UMLS, corresponding to 4,603 ICD-10 codes and to 10,078 MedDRA terms, included in 4,178 distinct CUIs (the same ICD-10 term can be mapped to several MedDRA term, and conversely). It means that approximately 32% of ICD-10 codes (national Clinical-Modifications are not considered) are mapped to MedDRA (PT or LLT level) in UMLS. The nature and proportion of SKOS predicates generated by the algorithm is described in the table below.

Table 1. Number of SKOS predicates generated by the algorithm and results of the expert evaluation.

Nb of SKOS predicates	exactMatch	narrowMatch	relatedMatch	broadMatch	Total
Generated	3769	3663	3645	569	11646
Evaluated sample	88	64	83	22	257
Validated by expert	82	25	13	10	130
95% confidence interval	$93.2^{\pm 5.3}$ %	$39.1^{\pm 12.0}$ %	$15.7^{\pm 7.8}$ %	$45.5^{\pm 20.8}$ %	50.6 %

Table 2. Example of SKOS properties generated by the algorithm.

ICD-10		MedDRA			SKOS match	evaluation
A00.9	Cholera, unspecified	10008634	LLT	Cholera, unspecified	exact	☑
		10008631	PT	Cholera	broad	☑
		10047397	LLT	Vibrio cholera gastroenteritis	related	☑
		10045658	LLT	Unspecified cholera	~~related~~	exact
A01.0	Typhoid fever	10045275	PT	Typhoid fever	exact	☑
		10045272	LLT	Typhoid	~~narrow~~	exact
		10039446	LLT	Salmonella typhi infection	~~narrow~~	exact
A02.0	Salmonella enteritis	10039433	LLT	Salmonella enteritis	exact	☑
		10039434	LLT	Salmonella gastroenteritis	related	☑
		10017914	PT	Gastroenteritis salmonella	~~broad~~	related
		10039443	LLT	Salmonella poisoning	related	☑

The results were evaluated on a random sample (257 mappings) by a medical terminology expert. All categories included, approximately 50% of the generated SKOS predicates proved to be correct, which is much better than chance level (equal to 25%, considering that only one of the four possible properties is correct in each case). skos:exactMatch properties, which are the most important for terminology alignment, were generated for 82% of ICD-10 terms for which mappings were available. Almost all are correct ($93.2^{\pm 5.3}$ %). The performances of the algorithm are less convincing for the other SKOS properties: only 15% of skos:relatedmatch and around 40% of skos:narrowMatch and skos:broadMatch predicates are correct.

3. Discussion

The algorithm described in this paper demonstrates satisfying performances, especially for the generation of skos:exactMatch properties, which suggests that automatic methods can be used to improve the quality of terminology mappings. The algorithm was tested on ICD-to-MedDRA mappings extracted from UMLS, but it can in principle be used for other terminologies and mapping sources.

Performances are more limited for other SKOS properties. This result is partly explained by the assumption that LLTs are more granular terms than PTs, which is not always verified. For example, *Cholera, unspecified* (LLT) is one subkind of *Cholera* (PT), but *Typhoid* (LLT) and *Salmonella typhi infection* (LLT) are synonyms of *Typhoid fever* (PT). In the former case, the algorithm (in that case rules 3 and 4 respectively) generates a correct SKOS predicate (namely skos:broadMatch), but not in the latter case: the correct mapping relation between the ICD-10 term *Typhoid fever* and the MedDRA LLTs *Typhoid* and *Salmonella typhi infection* is skos:exactMatch, because both refer to the same entity as the PT *Typhoid fever* (see Table 2).

An obvious way to improve the performances of our algorithm is consequently to distinguish between LLTs that are more granular terms and LLTs that are synonyms of PTs, to restrain the application of the rules described below to the former category, and to apply the same SKOS predicate when an ICD term is mapped to MedDRA terms that are synonyms. Considering that MedDRA 16.0 contains more than 70,000 LLTs, making such a distinction however requires a huge effort.

In addition, the current organization of MedDRA raises other issues. In some cases the semantic relation between LLTs and PTs is neither synonymy nor subsomption. For instance, *Salmonella enteritis* (LLT) is not a synonym or a specific kind of *Gastroenteritis salmonella* (PT), but refers to a related condition: both are kinds of Salmonella infections. One also finds cases where LLTs correspond not to *narrower* but to *broader* categories than PTs. This is for instance the case of the *Typhoid and paratyphoid fevers* (LLT), which is below *Typhoid fever* (PT) in MedDRA hierarchy, but corresponds to a broader category of disease. In ICD-10 and SNOMED-CT, *Typhoid fever* is under *Typhoid and paratyphoid fevers*.

Those elements suggest that the performances of algorithms such as the one presented here mainly depend on the quality of terminologies being aligned. Building terminologies with more systematic principles and more transparent hierarchical relations facilitates their alignment. Formal knowledge representation languages such as the Web Ontology language (OWL) can help in this task.

Acknowledgments The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement No. ICT-287800, SALUS Project. http://www.salusproject.eu/

References

[1] K.W. Fung, O. Bodenreider, Utilizing the UMLS for semantic mapping between terminologies, *AMIA Annual Symposium Proceedings* (2005), 266.

[2] K.W. Fung, O. Bodenreider, A.R. Aronson, W.T. Hole, S. Srinivasan, Combining lexical and semantic methods of inter-terminology mapping using the UMLS, *Stud Health Technol Inform* 129(2007), 605-609.

[3] International Classification of Diseases (ICD) Information Sheet, World Health Organization website. http://www.who.int/classifications/icd/factsheet/

[4] ICH guideline E2B (R2), Electronic transmission of individual case safety reports - Message specification (ICH ICSR DTD Version 2.1), Final Version 2.3, Document Revision February 1, 2001.

[5] G. Declerck, S. Hussain, Y. Parès, C. Daniel, M. Yuksel, A.A. Sinaci, G.B.L. Erturkmen, M.C. Jaulent. Semantic-sensitive extraction of EHR data to support ADE reporting, *SWAT4LS Proceedings* (2012).

[6] C.M. Hohl, A. Karpov, L. Reddekopp, J. Stausberg, ICD-10 codes used to identify adverse drug events in administrative data: a systematic review, *JAMIA* (2013), *doi:10.1136/amiajnl-2013-002116.*

[7] H. Sun, J. De Roo, M. Twagirumukiza, G. Mels, K. Depraetere, B. De Vloed, D. Colaert, Validation rules for assessing and improving SKOS mapping quality, *SWAT4LS Proceedings* (2013).

[8] A. Miles, S. Bechhofer, SKOS simple knowledge organization system reference, W3C Recommendation, 18 August 2009. http://www.w3.org/TR/2009/REC-skos-reference-20090818/

[9] ICH guideline E2B (R2), Electronic transmission of individual case safety reports - Message specification (ICH ICSR DTD Version 2.1), Final Version 2.3, Document Revision February 1, 2001.

[10] MedDRA 16.0. Term Selection: Points to Consider. ICH-Endorsed Guide, Release 4.5, April 2013.

[11] Merrill G. The MedDRA Paradox. AMIA Annu Symp Proc. 2008 6:470-4.

[12] MedDRA Introductory Guide Version 16.0. MSSO-DI-6003-16.0.0. March 2013.