Name:	Erik M. van Mulligen
Affiliation:	Erasmus University Medical Center, Dept. of Medical Informatics, Rotterdam, The Netherlands
Email:	e.vanmulligen@erasmusmc.nl
Presentation type (s):	Lightning talk

Annotating Dutch free-text patient records with OHDSI standard vocabulary concepts

Erik M. van Mulligen, PhD[1], Peter Rijnbeek, PhD[1], Jan A. Kors, PhD[1]
[1]Erasmus University Medical Center, Rotterdam, The Netherlands

Abstract

The OHDSI vocabularies only contain English terms, which cannot be used to annotate free-text electronic patient records in non-English languages. We explored the possibilities to use one of the OHDSI standard vocabularies, SNOMED-CT, for text-mining Dutch electronic patient records. We present the steps to automatically obtain a Dutch equivalent of the OHDSI vocabulary that can be used to mine standard concepts from unstructured text contained in the electronic patient records. We used different approaches to get a first impression of the coverage of this Dutch OHDSI vocabulary.

Introduction

Within our institute, we have converted the Dutch Integrated Primary Care Information (IPCI) database to the OMOP-CDM. The IPCI database contains a longitudinal collection of electronic patients records from 2.36 Million patients collected from 750 Dutch general practitioners (GPs), including a large corpus of unstructured data in the form of clinical notes. Part of the records also contain International Classification for Primary Care (ICPC) codes that were mapped to the CDM during the IPCI conversion process. The challenge is to effectively extract data from the clinical notes so that they can be used for observational research. The extracted concepts have to adhere to the OHDSI standard vocabularies when stored in the CDM. In this paper, we describe an approach to automatically translate English terms from one of the OHDSI standard vocabularies, SNOMED-CT, into Dutch, and assess the coverage in a sample of Dutch unstructured patient records. We also determine how many of the SNOMED codes that had been mapped from the ICPC codes in IPCI, had at least one Dutch term.

Approach

We extracted all SNOMED-CT concept identifiers from the OHDSI vocabulary (extracted August 7, 2017). These identifiers were used to find the corresponding concepts in the Unified Medical Language System (UMLS)[1] version 2016AB. We found 331,235 SNOMED concepts in

the OHDSI vocabulary, of which 261,944 (79%) could be mapped to UMLS. For each of the mapped concepts we retrieved any Dutch terms that were already available in the UMLS. This yielded 33,503 (13% of the mapped) concepts that had at least one Dutch term. See Table 1 for the results of these steps separated for the different domains.

Table 1. Results of mapping the SNOMED standard concepts to UMLS concepts and using UMLS to retrieve Dutch terms.

Domain	Mapped (%)	Translated (%)	Not mapped (%)
Condition	70,128 (91.9)	21,164 (30.2)	6,183 (8.1)
Measurement	13,405 (86.1)	1,572 (10.1)	2,172 (3.9)
Meas Value Operator	5 (100)	0 (0)	0 (0)
Meas Value	182 (96.8)	1 (0.5)	6 (3.2)
Device	14,764 (98.1)	248 (1.6)	294 (1.9)
Spec Disease Status	3 (100)	0 (0)	0 (0)
Unit	0 (0)	0 (0)	74 (100)
Spec Anatomic Site	25,338 (98.5)	1,255 (4.9)	387 (1.5)
Specimen	1,629 (96.7)	4 (0.2)	55 (3.3)
Relationship	151 (93.2)	24 (15.9)	11 (6.8)
Observation	96,315 (81.8)	6,676 (6.9)	21,363 (18.2)
Procedure	40,015 (88.4)	2,557 (6.4)	5231 (11.6)
Route	9 (42.9)	2 (22.2)	12 (57.1)
Overall	*261,944 (88.0)*	*33,503 (12.8)*	*35,788 (12.0)*

To get a first impression of the coverage, we determined how many ICPC codes in the structured data part of the IPCI records were mapped to SNOMED concepts that had at least one Dutch term. In Table 2 the results are shown. Almost all of the ICPC codes in IPCI can be mapped to UMLS. Only 18.4% of all unique ICPC codes appears to have a Dutch term associated, covering 50.0% of all instances.

Table 2. Per domain the number of IPCI codes used in Dutch electronic patient records that were mapped to UMLS identifiers and the number of concepts having at least one Dutch translation, counted per occurrence in the patient record and per unique code.

	All ICPC occurrences			Unique ICPC code		
Domain	**Mapped**	**Not mapped**	**Translated**	**Mapped**	**Not mapped**	**Translated**
Condition	39,794,188 (98.8)	482,994 (1.2)	6,101,476 (15.3)	765 (99.1)	7 (0.9)	103 (13.5)
Measurement	1,921 (3.4)	54,606 (96.6)	1,921 (100.0)	1 (50.0)	1 (50.0)	1 (100.0)
Observation	32,296,654 (99.9)	20,580 (0.1)	31,185,230 (96.6)	89 (97.8)	2 (2.2)	45 (50.6)
Procedure	4,238,533 (93.1)	314,557 (6.9)	841,804 (19.9)	33 (94.3)	2 (5.7)	14 (42.4)
Overall	*76,331,296 (98.9)*	*872,737 (1.1)*	*38,130,431 (50.0)*	*888 (98.7)*	*12 (1.3)*	*163 (18.4)*

Next, we analyzed how many of the SNOMED concepts with a Dutch term could be found in a

random sample of 100,000 lines from clinical notes in the IPCI database. We used the Dutch terms from UMLS. Terms that contained a semicolon were rewritten[2], e.g., "abnormaal; mictie" was changed into "mictie abnormaal". The resulting terms were used by our SolrTextTagger[3] based text-mining pipeline[4] to find concepts in notes. Table 3 shows per domain the number of concepts found. The 100,000 note lines contained 722,785 words. The high number of concepts can partly be attributed to homonyms. For example, the frequently occurring term "pols" (pulse) corresponded with four related but different concepts.

Table 3. Per type of concept an overview of the number of concepts found in the text of Dutch electronic patient records.

Domain	Concepts	Unique concepts
Condition	145,797	1,670
Spec Anatomic Site	29,450	284
Measurement	23,154	152
Specimen	437	2
Device	1562	43
Relationship	195	8
Observation	62,340	899
Procedure	15,726	197
Overall	*278,661*	*3,255*

Conclusion

Using UMLS as an intermediate step to translate the OHDSI vocabulary seems a reasonable first step. When mapping the codes from the electronic patient record to Dutch terms associated with the OHDSI standard concepts 50% have a translation. This figure cannot be used as an indication of how good concepts can be identified in a patient record text, but it could be used to get an impression of how many of the relevant concepts – many concepts from the OHDSI standard vocabulary are only rarely used – have a translation. In order to improve the mining of Dutch clinical notes we will improve the mapping of the OHDSI vocabulary to Dutch. We will use our experience with machine translation of vocabularies to extend the Dutch translation of the OHDSI vocabulary. In order to evaluate the quality of the text mining in Dutch, it is essential to have a manually annotated corpus of Dutch electronic patient record notes.

References

1. Bodenreider, Olivier. "The unified medical language system (UMLS): integrating biomedical terminology." *Nucleic acids research* 32. suppl_1 (2004): D267-D270.
2. Hettne, Kristina M., et al. "Rewriting and suppressing UMLS terms for improved biomedical term identification." *Journal of biomedical semantics* 1.1 (2010): 5.
3. David Smiley. 2013. Solr text tagger, text tagging with finite state transducers. https://github.com/OpenSextant/SolrTextTagger.
4. van Mulligen, Erik M., et al. "Erasmus MC at CLEF eHealth 2016: Concept Recognition and Coding in French Texts." *CLEF (Working Notes)*. 2016.