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

Erik M. van Mulligen, PhD[1], Peter R. Rijnbeek, PhD[1], Jan A. Kors, PhD[1], Johan van der Lei, PhD[1]

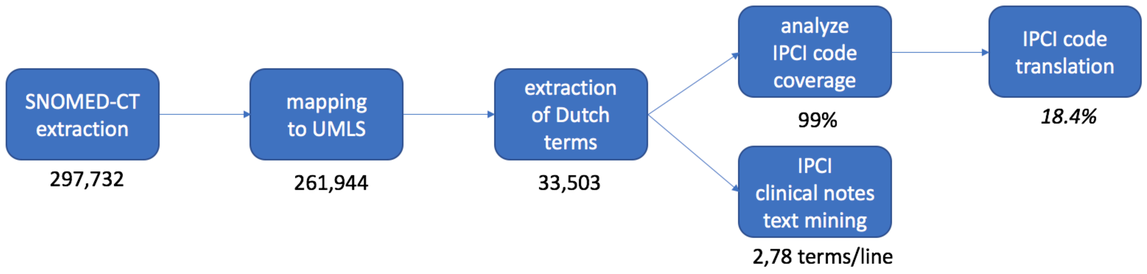
[1]Dept of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands

Background

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 for analyzing 3,6M patient records from 750 Dutch general practitioners available in the OMOP-CDM version of the Integrated Primary Care Information (ICPC) database.

Methods

From the OHDSI vocabulary unique SNOMED-CT concept identifiers were extracted. With the UMLS2016AB version these concepts were mapped to UMLS identifiers. For a subset of these UMLS identifiers at least one Dutch term was found. We further analyzed how this small set of Dutch terms covers the codes used in the IPCI database. The IPCI database uses ICPC codes for coding conditions, observations, procedures and measurements. These ICPC codes have almost all been mapped to UMLS. Only 18,4 of all unique codes have a Dutch term associated. We used the Dutch terms in a SolrTextTagger text mining pipeline to analyze 100K lines of clinical notes from IPCI. This yielded 278,661 terms recognized (2,78 term / line).



Results

The first table shows per domain and overall the figures for the mapping from SNOMED-CT to UMLS. The Translation column shows what percentage of the mapped terms contains at least one Dutch term. The second table shows how many of the ICPC codes were mapped to UMLS and how many of these have a Dutch translation. The third table shows the results of applying text mining to 100K lines from clinical notes in the IPCI database.

Conclusions

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. In order to improve the mining of Dutch clinical notes we will improve the mapping of the OHDSI vocabulary to Dutch, looking at those domains that have a low coverage first. 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.

Contact:e.vanmulligen@erasmusmc.nl

Results

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)

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

Domain	All terms (%)	Unique terms (%)
Condition	145,797 (52.3)	1,670 (51.3)
Spec Anatomic Site	29,450 (10.5)	284 (8.7)
Measurement	23,154 (8.3)	152 (4.6)
Specimen	437 (0.1)	2 (0.1)
Device	1562 (0.5)	43 (1.3)
Relationship	195 (0.0)	8 (0.2)
Observation	62,340 (22.4)	899 (27.6)
Procedure	15,726 (5.6)	197 (6.1)
Overall	278,661	3,255