

Identifying disease-centric subdomains in very large medical ontologies, a case-study on breast-cancer concepts in SNOMED.

Or: finding 2500 out of 300.000.

Krystyna Milian¹, Zharko Aleksovski², Richard Vdovjak², Annette ten Teije¹,
and Frank van Harmelen¹

¹ Vrije Universiteit Amsterdam, krystyna.milian@few.vu.nl

² Philips Research, zharko.aleksovski@philips.com

Abstract. Modern medical vocabularies can contain up to hundreds of thousands of terms. In any particular use-case only a small fraction of these will be needed. In this paper we first define two notions of a disease-centric subdomain of a large ontology. We then explore two methods for identifying disease-centric subdomains of such large medical vocabularies. The first method is based on lexically querying the ontology with an iteratively extended set of seed queries. The second method is based on manual mapping between terms from a medical guideline document and ontology concepts. Both methods include term-expansion over subsumption and equality relations. We use both methods to determine a breast-cancer-centric subdomain of the SNOMED ontology. Our experiments show that the two methods produce a considerable overlap, but they also yield a large degree of complementarity, with interesting differences between the sets of terms that they return. Analysis of the results reveals strengths and weaknesses of the different methods.

1 Introduction

Large medical ontologies such as SNOMED-CT³ contain hundreds of thousands of clinical concepts usually organized in a hierarchy and interconnected by domain specific relations, together representing the explicit semantic knowledge describing a medical field. Such knowledge can be of great help when developing intelligent clinical decision support systems that focus on reasoning about patient data within a certain disease domain. A disease-specific, richly annotated semantic subdomain is also an important element in the process of overcoming the frequent problem of lexical heterogeneity between the terms occurring in the patient data and those from the applicable clinical guidelines. However, identifying a *disease-centric subdomain* of a large medical ontology is not a trivial

³ <http://www.ihtsdo.org/snomed-ct/>

task. The relevant concepts are seldom to be found under one sub-branch of the ontology, instead they are usually scattered in various branches representing different facets of the domain coverage, e.g. clinical findings, procedures, anatomic regions, etc.

In this paper we describe a study on the identification of SNOMED concepts related to breast-cancer. We compare results of two different methods: (i) The *seed query method* from [1] was used for extraction of concepts that are unique to breast-cancer. (ii) The so-called *guideline-based method*, consisting of a manual mapping between SNOMED concepts and the important terms from the Dutch national breast-cancer guidelines, was used for the identification of those concepts that are relevant with respect to breast-cancer.

Our experiments show that the two methods produce a considerable overlap, but they also yield a large degree of complementarity, with interesting differences between the sets of terms that they return. The size of the identified subdomains is considerably smaller than that of the whole medical ontology (between 0.1%-1%), making the reasoning as well as the maintenance task of such subdomain much more feasible.

The paper is structured as follows: Section 2 introduces different notions of relevancy in subdomains of a medical ontology, and puts forward the main hypothesis of the paper. Section 3 and 4 introduce our two different subdomain-selection methods: the seed query method in section 3 and the guideline-based method in section 4. Section 5 compares and analyses the results. Section 6 summarizes the findings and presents the concluding remarks.

2 Two types of disease-centric subdomains

Before investigating methods for identifying disease-centric subdomains from a large ontology, we must first define what we mean by such a subdomain. We will briefly look at other work that aims at identifying subdomains in large ontologies, and will then set out our own definitions.

Related work: Existing methods in the literature often rely on an a priori modularization of the vocabularies. These are typically based on some notion of semantic distance, or on the connectivity-graph of the ontology [3,8]. Instead, our methods are not based on any *a priori* modularization of the ontology, but they identify subdomains that are specific for any particular use of a vocabulary.

Definitions: We distinguish two kinds of disease-centric subdomains, namely *relevant subdomains* and *key subdomains*, which consist of relevant terms and key terms respectively. The notions of “relevant terms” and “key terms” are each defined as follows:

Relevant Terms: A term T is a *relevant term* for a disease D if it is contained in a source which influences decisions on the diagnosis or treatment of D.

An example of a term that is relevant to breast-cancer is “pregnancy”: data-sources about breast-cancer (such as guidelines, patient-records, textbooks, etc) often contain the term “pregnancy” because certain treatments (e.g. chemotherapies) are ruled out for pregnant women.

However, the converse is not the case: not any document containing the term “pregnancy” is likely to be about breast-cancer. To capture this, we define a second notion:

Key terms: A term T is a *key term*⁴ for a disease D if the occurrence of T in a datasource S means that S is surely about D

An example is the term “malignant neoplasm of breast”. Any key term is of course a relevant term, but not vice versa.

Hypothesis: Our hypothesis is that the seed query method (described in section 3), when seeded properly, will identify only key concepts, while the manual guideline-based method (described in section 4) will identify relevant concepts. From the above definitions, this hypothesis also implies that the seed query results should be contained in the guideline-based results.

Choice of dataset: In this paper, we focus on breast-cancer as our clinical domain both because of its prevalence and the highly progressed state-of-the-art in diagnoses and treatment, which is expected to involve a relatively rich vocabulary and thus presents an interesting use-case. We concentrate on SNOMED-CT as our main ontology, mainly because of its high adoption and a broad clinical coverage, containing more than 300.000 concepts. Besides applying both methods to the breast-cancer domain in SNOMED, we also apply the seed query method to three other very large ontologies to verify the consistency of our results. Also applying the manual guideline-method to all these ontologies would have been prohibitively expensive.

3 Seed query method to find key terms

Method The seed query method, originally published in [1], is a combination of a lexical and a structural approach. It takes a list of terms (the so-called “seed queries”), which serve as prior knowledge, to find an initial set of breast-cancer concepts through lexical mapping to the concepts in the ontology. This set is then expanded through the hierarchical structure of the ontology, and through the semantic network of UMLS. Given a set of seed queries, the process is completely automatic, ensuring repeatability of the extraction. It also allows for gradual improvement by adjusting the initial set of seed queries.

In more detail, the seed query method proceeds in three steps: (i) *Query matching* which uses the concept’s names, (ii) *Subconcept expansion* based on the hierarchical structure of the ontologies, and (iii) *UMLS expansion* which uses

⁴ “key” is inspired by the database notion of the same name

the UMLS metathesaurus. The three steps in this method are incremental, each step produces set of concepts which is passed as input to the next step. The third step produces the final result of the method.

Query matching uses a list of seed queries to find concepts from the subdomain by trying to lexically match the queries to each concept from the ontology. The lexical match was not sensitive to letter capitalization, and in addition, Porter's stemmer algorithm [6] was used to normalize the words before comparison. Such queries consist of keywords or combinations of keywords which are specific to the subdomain, and when a concept lexically matches to some of these queries, it can be considered part of the subdomain. The algorithm for query matching is shown in Figure 1.

Subconcept expansion expands the set of concepts produced in the first step by including their subconcepts. Each ontology generally organizes the concepts in a hierarchy through IS-A relations among them (e.g. Breast-cancer IS-A Cancer). These relations were used to find all the subconcepts to the concepts found in the first step. This process was done exhaustively, transitively adding the subconcepts of the newly found concepts as well, until no new concepts could be added. The algorithm for subconcept expansion is shown in Figure 2.

UMLS *expansion* further increases the set produced in the second step in the following way: if a concept that was found in the first two steps exists in another ontology as well, according to UMLS, and it was not found in the first two steps, then it is added in the appropriate set of that other ontology. UMLS assigns a unique identifier to every concept from every ontology integrated in it. If two concepts have the same identifier then they are equivalent by meaning. The algorithm for UMLS expansion is given in Figure 3.

Results The breastcancer-centric subdomain of SNOMED (containing only key concepts for breast-cancer) was extracted using the method described above.

We seeded the method with a hand-crafted list of breast-cancer seed queries, shown in Table 1. After starting with a small number of key terms, and iteratively adding seeds, we observed that after a small number of terms the results stabilise, and no longer grow when adding further key terms as seeds. This process has up to now been informal, and would merit a more detailed study in its own right.

Besides SNOMED, the method was applied on three other ontologies: NCI⁵ - a vocabulary for annotating medical documents primarily cancer related, MeSH⁶ - a vocabulary for scientific literature annotation and ICD10⁷ - a classification of diseases. The ontologies were used as extracted from the UMLS 2008AA version.

The results of applying the seed query method are shown in Table 2. The table shows that only a fraction of the entire ontology (much less than 1%) are key terms for a disease such as breast-cancer. It also shows that most of the results are actually found in the first phase. This is reasonable: most of the concepts are very specialized and are hence leaves in the ontologies. Finally, it is

⁵ <http://nciterms.nci.nih.gov>

⁶ <http://www.nlm.nih.gov/mesh>

⁷ <http://www.ahima.org/icd10>

The resulting set of matched concepts is empty in the beginning

```

1 subdomain :=  $\emptyset$ 
  Lexically matching the concepts from the ontology to the query list
2 for each query  $Q \in$  list of queries do
3   for each concept  $C \in \mathcal{C}^{\text{ONT}}$  do
4     if LEXICALMATCH( $C, Q$ ) and  $C \notin$  subdomain then
5       subdomain  $\leftarrow C$ 

```

Fig. 1. Step one: Query matching.

Add all the subconcepts to the concepts in subdomain

```

1 while changes are possible repeat
2   for each concept  $X \in$  subdomain do
3     for each concept  $Y \in \mathcal{C}^{\text{ONT}}$  do
4       if  $Y \subseteq X$  and  $Y \notin$  subdomain then
5         subdomain  $\leftarrow Y$ 

```

Fig. 2. Step two: subconcept-based expansion.

Expanding each of 4 result sets through UMLS

```

1 for any two ontologies  $\text{ONT}_p, \text{ONT}_q \in \{\text{SNOMED}, \text{NCI}, \text{MeSH}, \text{ICD10}\}$  do
2   for each concept  $X \in \text{subdomain}_p$  do
3     for each concept  $Y \in \text{subdomain}_q$  do
4       if UMLS :  $X \equiv Y$  and  $Y \notin \text{subdomain}_q$  then
5          $\text{subdomain}_p \leftarrow Y$ 

```

Fig. 3. Step three: UMLS-based expansion.

interesting to see that the most specialised ontology (the oncology-specific NCI) has the highest hit-rate of key terms, and the most general and wide ranging ontologies (SNOMED and ICD10) have the lowest hit-rates.

4 Manual mapping of guidelines to find relevant terms

We used the official guidelines for the treatment of breast-cancer as a source of information to identify the relevant breastcancer-centric subdomain. Medical guidelines describe recommendations and conclusions regarding proper treatment based on scientific evidence. They aim to reduce the growing gap between knowledge and the actual practice. We used in our research guidelines developed by the joint initiative of the Dutch Institute for Healthcare Improvement (CBO) [2].

From formalised models of the guideline [4] we extracted the names of all treatment plans, as well as all parameters describing patient data and their possible values in case of enumerated types. The parameters either specify plan preconditions and intentions or data that can be requested from external sources during guidelines execution.

Table 1. Seed queries used to extract the breast-cancer subdomain.

1. Breast cancer
2. Breast carcinoma
3. Microcalcification
4. Mammary carcinoma
5. Lobular carcinoma
6. Ductal carcinoma
7. Mastectomy
8. Paget breast
9. HER2/neu
10. HER-2
11. BRCA

Table 2. Results of applying the seed query method on the four ontologies: incremental results are reported after each step (full method = after step 3)

Ontology	size of ontology	number of concepts extracted			% of full ontology
		after step 1	after step 2	after step 3	
SNOMED	308,677	198	271	279	0.09%
NCI	62,969	358	388	399	0.63%
MeSH	282,425	105	120	129	0.05%
ICD10	11,529	5	5	12	0.10%

Practical experiences The main challenges of mapping terms extracted from the guidelines to SNOMED concepts were searching among the hundreds of thousands of SNOMED concepts for the equivalences. Mapping required understanding the meaning of terms used in the guidelines and knowing the exact context where they were used. After the initial mappings were identified, we consulted with our clinical expert and made adjustments where necessary. Below we illustrate some of the difficulties which we encountered.

In many cases guidelines and SNOMED use different terminology to express the same information. 'Axillary-node-dissection-proper' used in the guidelines and 'Excision of axillary lymph node' defined in SNOMED are an example of such case. Finding corresponding terms was done using key words or using synonyms found in medical dictionaries. In cases where both approaches failed, we checked the context in the guidelines or looked for an explanation of terms in other resources. This applied in the case of abbreviations as well as full phrases.

On the other hand finding an exact lexical match can be sometimes misleading. Such a situation was encountered when the plan 'Mastectomy' was analyzed. In the guidelines it covers the plan 'Mastectomy-proper' and also other procedures such as 'Radiotherapy-chest-wall' and 'Breast-reconstruction'. Hence the plan 'Mastectomy-proper' rather than 'Mastectomy' should be mapped to the SNOMED term 'Mastectomy'. Therefore knowing the context was necessary.

Differences in granularity and abstraction level caused most of the missing matches. This issue appears mostly in the case of multiterms expressions, which are commonly used in the guidelines. Examples of such compound terms are therapy + drug e.g. anthracycline-chemotherapy-manual, or therapy + drug + number of repetition e.g. six-courses-anthracycline-chemotherapy. Multiterms expressions are also used to define the intentions of therapies, for example 'elimination-distant-metastases'. Such specific concepts turned out to be very unlikely to be found in SNOMED ontology.

In a few cases even the large SNOMED ontology is not expanded enough yet. For example, SNOMED contains no concept corresponding to the parameter 'patient-preferes-bct', describing the patients preference of breast conserving treatment over mastectomy.

All these points above show that the method of obtaining relevant subdomains by mapping from guidelines is essentially a manual operation that cannot easily be automated. Our early results with such automated procedures ([7]) also corroborate this.

Results of the manual mapping We found around 60 exact matches (matches with the same meaning but not necessarily the same name) out of all 150 parameters extracted from the guidelines. In the case of treatment procedures, we found around 40 exact matches out 170 procedures, and 40 matches, where SNOMED terms have a close but more general meaning. The missing matches are caused by the reasons mentioned above.

Results of the expansion steps In section 3, seed queries were used for the lexically querying for matching terms. In the guideline-based method, this step is performed more semantically, namely by manually mapping the parameters and procedures of the guideline. In both cases, this first step is followed by subconcept-based expansion (transitively including all subsuming concepts, fig. 2) and UMLS expansion (using UMLS to include equivalent concepts, fig. 3).

Applying these two expansion steps to the results of the first manual mapping step resulted in an expansion from 140 to 2250 terms. The two expansion steps have a much bigger impact after the manual mapping (from 140 to 2250) than they have after the first step in the seed query method (from 198 to 279). This difference can be explained by the fact that the first step in the seed query method returns mostly very specific SNOMED concepts that have very few subconcepts, while the manual mapping also yielded concepts higher in the SNOMED hierarchy.

However, also in the manual mapping case, the breastcancer-centric subdomain is again a very small fraction of the entire ontology, namely 0.73 % of the full ontology (308.677 terms).

5 Evaluation of the two methods

Our two methods for identifying breastcancer-centric subdomains provided different results. The manual guideline-method found 2250 terms, against 279 terms

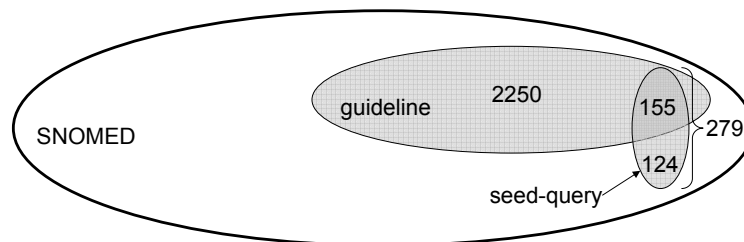


Fig. 4. Breast cancer subdomains identified using different approaches.

found by the seed query method. Of these 279 terms, 155 are also found by the guideline-method. The inclusion relations are summarised in figure 4.

Unsurprisingly, all 2250 terms found by the guideline-method are indeed relevant terms for the breastcancer-centric subdomain, in other words this method has a high precision. This is unsurprising because all terms are either direct mappings from parameters or procedures in the recommendations of a national breast-cancer guideline, or are subconcepts of these terms.

More interesting, manual inspection of the 279 seed query results shows that this method has a near perfect precision (i.e. all the terms it finds are indeed key-terms for the breastcancer-centric subdomain). This confirms the main hypothesis put forward in section 2.

The figure also shows that besides its high precision (finding only key terms), the seed query method has a rather low recall: it finds less than 10% of the terms found by guideline-method. This is to be expected since the seed query method is tuned to find only key-terms (instead of finding all relevant terms). However, inspection of the 2095 terms that are only found by the guideline-method reveals that there are quite a few key terms still contained in that set. Hence, even when counting only key terms, the seed query method has no perfect recall. Examples of obvious terms that we found missing are “Breast surgical margin involved by tumor”, very detailed terms such “Metastasis in internal mammary lymph nodes with microscopic disease detected by sentinel lymph node dissection but not clinically apparent” and quite a few others.

Finally, and contrary to our prediction, the seed query results are not a subset of the results from the guideline-method. In fact, well over 40% of all seed query results (124) are not found by the guideline-method. Inspecting this set yielded the following explanations for this falsification of our hypothesis:

Guideline does not cover diagnostic terms: The biggest part of terms in this group describe breast neoplasm in general, e.g. ‘Carcinoma in situ of female breast’. The guidelines are focused on recommendation for treatment of already diagnosed breast-cancer, which is malignant. Benign neoplasm is not broadly discussed, since such terms would be rather covered by diagnosis guidelines.

Only the guideline recommendations were used: Some of those terms are connected with breast-cancer but are not included in the recommendations, the

only part of the guidelines which was formalized. For example recommendation do not mention treatment procedures for male breast-cancer, whereas terms like 'Carcinoma in situ of male breast' or 'Carcinoma in situ of areola of male breast' were identified by seed queries.

Recent medical insights are not included in the guideline: Among this group appear also terms about gene findings like 'BRCA1 gene mutation positive'. These are only recently taken into account during breast-cancer treatment. Gene findings are covered by the new version of the guideline [5].

The guideline does not mention procedures that vary between hospitals: In The Netherlands, some hospitals employ special oncology nurses for home care of patients, others don't. The national guideline does not discuss procedures for which there is an accepted high local variance between hospitals.

Between them, these reasons would remove a substantial part of the outlying 124 concepts, although we are currently not able to determine the exact number.

6 Summary and Conclusions

Summary Medical vocabularies are typically very large, containing up to hundreds of thousands of concepts. However, for any particular usage of such vocabularies only a small fraction of the concepts will be needed. In our example use-case, the breastcancer-centric subdomain of SNOMED is at most 1% of all concepts in the ontology. This gives urgency to the question of how to find such relevant subsets of terms from potentially very large vocabularies.

In this paper, we have investigated two methods for identifying such relevant concepts. Our first method consisted of manually identifying a number of seed queries, and performing a lexical search for all concepts whose lexical labels contain any of the seed queries as a substring. All of the resulting concepts and their subconcepts are then considered as relevant for the subdomain characterised by the seed queries followed by the expansion phases. Our second method consisted of manually identifying all SNOMED terms that appeared as a parameter or procedure in the recommendations of the Dutch national guideline for the treatment of breast-cancer, again followed by the two expansion phases.

These methods differ from other approaches for the identification of relevant subvocabularies that are available in the literature: they are not based on any *a priori* modularization of the ontology, but instead select sets of concepts that are specific for a particular use of a vocabulary.

Conclusions Our findings indicate that:

- the breastcancer-centric subdomain is indeed only a fraction (< 1%) of all terms in SNOMED
- the seed query method has a high precision, returning only key concepts
- the seed query method has a low recall for returning relevant terms
- the guideline-method has a higher recall for relevant terms while still having a high precision for relevant (but possibly non-key) terms.

- contrary to our prediction, not all key-terms are found by the guideline-method. Close inspection yielded a number of reasons why this is the case in our experiment:
 - the guideline covers only procedures for treatment, hence misses diagnostic concepts
 - we extracted our concepts only from the recommendations in the guideline, hence missing those concepts that only appear in the background information
 - the guideline does not mention procedures that vary between hospitals
 - the guideline is not yet updated with recent insights about molecular and genetic markers for breast-cancer, while these concepts did appear in our seed queries

Future Work In future work, the validity of our conclusions should be tested by running these experiments on other subdomains (e.g. different diseases), and possibly using other methods to obtain a "gold standard" (our gold standard was obtained by manual extraction of all concepts from a national treatment guideline).

Similarly, it would be interesting to apply the guideline-method to other documents such as patient-records to see if that would yield a very different set of terms.

More insight should be obtained in the correct choice for the seed terms, since obviously the method is sensitive to this. The apparent fixed-point behaviour of this method deserves further investigation, for example on the degree of sensitivity to the initial set of query-terms.

References

1. Z. Aleksovski and R. Vdovjak. Overlap of selected ontologies in the context of the breast cancer domain. In *Proceedings of SIIM 2009*, 2009.
2. CBO. *Guideline for the Treatment of Breast Carcinoma*. van Zuiden, 2002. PMID: 12474555.
3. B. Cuenca Grau, I. Horrocks, Y. Kazakov, and U. Sattler. Just the right amount: extracting modules from ontologies. In *Proceedings of WWW*, pages 717–726, 2007.
4. M. Marcos, J. C. Galan, B. Martinez, C. Polo, A. Seyfang, S. Miksch, R. Serban, A. ten Teije, F. van Harmelen, K. Rosenbrand, J. Wittenberg, J. van Croonenborg, P. Lucas, and A. Hommersom. Protocure ii deliverable d2.2bcd: Models of selected guideline in intermediate, asbru and kiv representations. Technical report, www.protocure.org, 2005.
5. Nationaal Borstkankeroverleg Nederland, Kwaliteitsinstituut voor de Gezondheidszorg CBO, and Vereniging van Integrale Kankercentra. Richtlijn mammacarcinoom 2008, 2008.
6. M. F. Porter. *An algorithm for suffix stripping*, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.
7. R. Serban and A. ten Teije. Exploiting thesauri knowledge in medical guideline formalization. *Methods of Information in Medicine*, 2009. To appear.
8. H. Stuckenschmidt and M. Klein. Structure-based partitioning of large concept hierarchies. In *International Semantic Web Conference*, pages 289–303, 2004.