

# Identifying Disease-centric Subdomains in Very Large Medical Ontologies, a Case-study on Breast-cancer Concepts in SNOMED

Krystyna Milian<sup>a</sup>      Zharko Aleksovski<sup>b</sup>      Richard Vdovjak<sup>b</sup>  
Annette ten Teije<sup>a</sup>      Frank van Harmelen<sup>a</sup>

<sup>a</sup> *Vrije Universiteit Amsterdam, krystyna.milian@cs.vu.nl*

<sup>b</sup> *Philips Research, zharko.aleksovski@philips.com*

The full version of this paper appeared in: Workshop on Knowledge Representation for Health-Care (KR4HC 2009) in conjunction with the 12th Conference on Artificial Intelligence in Medicine (AIME'09), Italy, July 2009.

## 1 Introduction

Large medical ontologies such as SNOMED-CT<sup>1</sup> 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. Identifying a *disease-centric subdomain* of such a large medical ontology is not a trivial 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 the full paper we describe a study on the identification of SNOMED concepts related to breast cancer. We compare the results of two different methods: (i) the *seed query method* from [1] was used for extraction of concepts that are unique to breast cancer, and (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 for identifying relevant concepts 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, and that they identify a subdomain which is considerably smaller than that of the whole medical ontology (between 0.1%-1%).

## 2 Two types of disease-centric subdomains

We distinguish two kinds of disease-centric subdomains, namely *relevant subdomains* and *key subdomains*, which consist of relevant terms and key terms respectively.

**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”: datasources about breast-cancer (such as guidelines, patient-records, etc.) often contain the term “pregnancy” because certain treatments (e.g. chemotherapies) are ruled out for pregnant women.

**Key terms:** A term T is a *key term* 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”.

---

<sup>1</sup><http://www.ihtsdo.org/snomed-ct/>

Any key term is of course a relevant term, but not vice versa.

**Hypothesis:** Our hypothesis is that the seed query method, when seeded properly, will identify only key concepts, while the manual guideline-based method 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.

### 3 Methods for identifying disease-centric subdomains

**Seed query method to find key terms** 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.

**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. From formalised models of the guideline 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.

Both methods differs 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.

### 4 Findings

We have investigated the two methods and 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 breastcancer, while these concepts did appear in our seed-queries

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 subdomain 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 a subdomain much more feasible.

### References

- [1] Zharko Aleksovski and Richard Vdovjak. Overlap of selected ontologies in the context of the breast cancer domain. In *Proceedings of Society for Imaging Informatics in Medicine (SIIM 2009)*, 2009.