Evaluating the Coverage of three Controlled Health Vocabularies with Focus on Findings, Signs & Symptoms

Dimitrios Kokkinakis
Dep. of Swedish (Språkbanken) & Centre for Language Technology (CLT)
University of Gothenburg, Sweden
dimitrios.kokkinakis@svenska.gu.se

Abstract

The medical domain is blessed with a magnitude of terminological resources of various characteristics, sizes, structure, depth and breadth of descriptive power, granularity etc. In this domain a particularly interesting and difficult entity type are signs, symptoms and findings which to a large extend are expressed in a periphrastic manner, sometimes by the use of figurative or metaphorical language, or contextualized using a wealth of vague variant expressions. We hypothesize therefore that no major official terminology source alone can accommodate for the variation and complexity present in real text data, such as electronic medical records, notes or health related documents. In this paper we evaluate the content of the three largest medical control vocabularies available for Swedish on extracted reference symptom lists and initiate a discussion on how we should proceed in order to accommodate for increased coverage on similar genres.

1 Introduction

The medical domain is blessed with a magnitude of terminological resources of various characteristics, sizes, structure, depth and breadth of descriptive power, granularity etc. This paper deals with a first attempt to investigate, understand and in the future harmonize large medical terminological resources with focus on a particular interest and difficult to describe type of terms, namely *signs*, *symptoms*, *findings* and other *symptom-based phenotypes*. We hypothesize that no major official terminological source alone can accommodate for the variation and complexity for such terms present in real text data. Preliminary experiments indicate that to a great extend signs, symptoms and findings are expressed in a periphrastic manner, sometimes by the use of figurative or metaphorical language, or contextualized using a wealth of vague variant expressions. However these characteristics seem to vary depending on the type of data examined. In this paper we evaluate the content of the three largest medical control vocabularies available for Swedish on extracted reference symptom lists and initiate a discussion on how we should proceed in order to accommodate for increased coverage.

The followed approach can be seen as exploratory in which we believe to yield insights into the nature of symptom contextualization in order to be able to enhance our knowledge of communicative events in various healthcare settings. This study is initiated in the context of a recently started project, entitled *Interpretation and understanding of functional symptoms in primary health care*. The main research goal of the project is to study health care interactions with patients suffering from *Functional Somatic Syndromes* (FSS). Relevant research has showed that the care actions taken within primary health care are unsuccessful in the purpose to reduce the patients' suffering. The project's hypothesis is that the interaction in patient/care provider encounters is dysfunctional because of diverging perspectives and interpretation frames. This is resulting in lack of understanding and explanation of the patients' symptoms, leading to unsatisfaction and frustration among patients as well as care providers. One of the project's strand of research activities is on investigating how symptom mentions are expressed and how successful automated means are for capturing symptom descriptions both on collected written (patient records) and transcribed material (patient/nurse and patient/doctor encounters).

2 Background

The medical domain is particularly well endowed with sources of terminology, but there is also a large body of work with emphasis on methods for building required terminological knowledge bases automatically or semi-automatically from

textual sources. This is guided by the assumption that even though substantial term lists are available, automated methods have the benefit of being able to discover new variant terms, acronyms etc. and add them to existing lists (*cf.* Grishman *et al.*, 2002; Krauthammer & Nenadic, 2004; Tsujii & Ananiadou, 2005). Consequently, evaluation of terminologies in various subdomains has shown that there is a long way to go in order to achieve complete coverage. For instance, Langlotz & Caldwell (2005) discuss that no lexicon achieved greater than 50% completeness for any test set of imaging terms and that no single lexicon was sufficiently complete to allow comprehensive indexing, search, and retrieval of radiology report information.

Our work is also inspired to a certain degree by Unified Medical Language System (UMLS®; Kohler, 2008) since it would have been desirable in the future to have such comprehensive platform for e.g. Swedish. UMLS facilitates the development of computer systems that behave as if they *understand* the meaning of the language of biomedicine and health. The main purpose of the UMLS is to facilitate conversion of terms from one controlled medical vocabulary to another. UMLS consist of three knowledge sources, the *Metathesaurus*®, the *Semantic Network*, and the *SPECIALIST Lexicon*. The Metathesaurus forms the base of the UMLS and comprises several million concept names, all of which stem from the over 100 incorporated controlled vocabularies and classification systems. Some examples of the incorporated controlled vocabularies are ICD-10, MeSH, SNOMED CT, DSM-IV, LOINC and the Gene Ontology.

3 Controlled vocabularies (for Swedish)

3.1 Symptoms vs. Signs

In general terms, a symptom is a manifestation of a disease, indicating the nature of the disease, which is noticed by the patient; in this respect symptoms are *subjective* by nature. This is usually contrasted to signs which are observed by a medical practitioner and are thus *objective* measures by nature. Sometimes the context is important in order to distinguish one from the other, while often the distinction is blurred.

3.2 MeSH, SNOMED CT & KSH97/ICD-10

The Medical Subject Headings (MeSH) under the hierarchy C (*Disorders*) incorporates the subhierarchy C23 (*Pathological Conditions, Signs and Symptoms*) which includes abnormal anatomical or physiological conditions and objective or subjective manifestations of disease, not classified as disease or syndromes. The Swedish MeSH (edition 2006) includes 880 term entries in C23 which we also use in the current study, examples include *smärta* 'pain', *svullnad* 'edema' and *nysning* 'sneezing'.

The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) is a systematically organized computer processable collection of medical terminology covering most areas of clinical information. A relevant top level hierarchy in SNOMED CT is *finding*. The Swedish version of SNOMED CT (first release of April 2010) includes 32 911 findings, such as *brännande känsla* 'burning feeling' (90673000), *undernärd* 'malnourished' (248325000) and *kronisk hosta* 'chronic cough' (68154008).

Finally, the International Statistical Classification of Diseases and Related Health Problems (ICD) contains a listing of chapters one of which, Chapter XVIII, *Symptoms, signs and abnormal clinical and laboratory findings*, is relevant for this study. XVIII contains 532 terms, examples include *onormal hjärtrytm* 'abnormal heart rhythm' (R00), *dysuri* 'dysuria' (R30.0) and *dåliga matvanor* 'unhealthy nutrition habits' (R63.3). The Swedish translation of ICD is based on the Classification of Diseases 1997 (KSH97) and a systematic list that was released in September 1996. KSH97 (ICD-10) was recently replaced by ICD-10-SE (January 2011). In this study we use the older version.

4 Material and Method

There are several health related portals on the internet that provide a rather thorough description of diseases, their symptoms, etiology, treatment etc. The data sources of the symptoms' encoding used for the empirical evaluation were extracted from three popular health portals. The first site is intended for professional users, i.e. medical doctors <http://www.praktiskmedicin.com> the second and third are intended for laymen <http://www.netdoktor.se> & <http://www.1177.se>.

Fifteen randomly selected disease description pages were visited from each portal (Appendix A1). The symptoms' discussion parts for each disease was transferred to an external file, tokenized and automatically annotated with the three terminologies. The total number of manually identified symptoms was 552 (475 unique).

5 Evaluation

For the evaluation of the existing terminologies we chose a pragmatic approach as previously outlined, since available gold standards for such evaluation do not exist. A quantitative and qualitative analyses of the results are shown in table 1. Qualitative analysis in this context implies a thorough, manual examination of each annotated symptom description mention. A process that allows us to get a clearer picture on how symptom descriptions are formulated in text, what the limitations of the terminologies are and whether there is a need of harmonization of the terminologies and the gains we can expect. Moreover, it became apparent that enhancement with other mechanisms, such as extensive inclusion of variant forms (if available) or links to laymen vocabularies is necessary in order to enable a highly accurate and sufficient coverage of the textual content.

5.1 A Reference List

For ease of evaluation we chose to manually produce three reference lists, one from each site, a part of the accumulative term list is given in-Appendix A2. The quality of the controlled vocabularies, with respect to coverage, was evaluated in terms of i) the number of exact matches of text mentions; ii) the number of exact matches of text mentions after semiautomatic enhancement of the terminologies with various transformed variants (*cf.* Kokkinakis, 2009); iii) partial matches after the vocabulary enhancements and iv) the number of non-match after the vocabulary enhancements. The average symptom is 2,46 words long. Table 1 summarizes the results.

	ND	**1177**	**PM**
SNOMED	31,9%	34,4%	**34,9%**
MeSH	22,4%	24%	**26,8%**
ICD-10	3,2%	**4,3**%	4,1%
SNOMED+	38,1%	51,2%	45.6%
MeSH+	29,4%	32,8%	34,4%
ICD-10+	5,1%	6%	4,9%
No match SNOMED+	6,2%	19,2	13,9%
No match MeSH+	9,5%	13,6%	15%
No match ICD-10+	85,2%	85,4%	86,8%
Partial SNOMED+	55,6%	29,6%	40,3%
Partial MeSH+	60,9%	53,6%	50,5%
Partial ICD-10+	9,7%	8,6%	8,3%

Table 1: Evaluation results based on three samples (ND: *NetDoktor* and PM: *Praktisk Medicin* and 1177: *1177.se*) without/with vocabulary extensions (variations) the latter designated by the *plus* sign.

In the table above *Partial* implies that the obtained annotation is not complete. Sometime partial matching is sufficient in order to grasp the meaning of a text sequence such as in the example *hörselnedsättning på ena örat* 'hearing loss in one ear' in which both *hörselnedsättning* (C23.888.592.763.393.341) and *örat* (A01.456. 313;A09.246) have been recognised by MeSH but not the whole composite term. In other cases partial annotation is insufficient to capture the proper meaning such as in the case of *rasslande ljud i bröstkorgen* 'rattling sound coming from the chest cavity' in which only *bröstkorgen* could be matched.

6 Discussion

The initial findings of this study suggest that in combination the three resources have the potential to adequately represent a large number of the terms required to describe symptoms. All three together provide substantially more exact matches than any individual vocabulary in the set, although SNOMED CT gives the far better results. This is a natural consequence since its content is far more extensive and nuanced than both MeSH and ICD-10 together.

A problem faced with our approach is the fact that it is hard to determine whether potential missed terms (i.e. unmatched) were truly "absent" from the vocabularies or there might have been synonyms/variants in the resources that could not be identified despite the use of a large number of generated variant forms and near synonyms. Another important issues is the difficulty, in some cases, to differentiate between findings/symptoms and disorders/diseases. Although there is a separation in the three resources, sometimes fuzzy, as indicated in the MeSH-SNOMED distinction, in which a number of findings according to SNOMED were labeled with other hierachies in MeSH, such as *irritabilitet* 'irritable mood' which is found with the label "F01.470.047.110" which belongs to the *Psychiatry and Psychology* hierarchy; or *högt blodtryck* 'high blood pressure' which is found with the label "C14.907.489" which belongs to the *Cardiovascular Diseases* subhierarchy. However, these cases were marked as correct.

While an absent synonym can be remedied by simply adding a surface form, a missing concept represents a more significant absence but we could not identify such cases *cf.* the discussion by Wasserman & Wang (2003). There were a small number of lexical ambiguities (homographs) such as the phrase *sena skeden* litt: 'late stages' for which the SNOMED returned an an-

notation for *sena* ‘tendon’ (body structure) and *skeden* ‘the spoon’ (physical object); obviously both annotations are wrong in this context. Although the sample is not spontaneous language a number of metaphoric and figurative language expressions could still be found, such as *brännande smärtor* ‘burning paint, *bubblig i magen* ‘bubbly in the stomach’, *månansikte* ‘moon-face’, *buffelpuckel* ‘buffalo hump’, *motorisk klumpighet* ‘motor clumsiness’ and *produktiv hosta* ‘productive cough’. Finally, an issue that needs attention is various types of coordinations that need to be resolved in order to increase coverage, such as *minnes- och koncentrationsstörning* ‘memory and concentration disturbance’ and *fingrarnas ytter- och mellanleder* ‘fingers outer and middle joints’ and which may be resolved as *minnesstörning* & *koncentrationsstörning and fingrarnas ytterleder* & *fingrarnas mellanleder*.

7 Conclusions

Term matching in new subdomains of medicine is likely to identify further omissions highlighting the importance of a responsive updating process (Brown & Odusanya 2001). In the near future we intend to make detail analyses of other types of data, patient records and transcribed data, which will shed more light to whether controlled vocabularies can capture the patients' contextualization of symptoms, which is the main focus of this initiated activity. For future work we also intend to investigate whether partial or uncaptured symptom/finding-like terms are parts of disease/disorder descriptions. There might be other sources of lexical/terminological knowledge that might have been useful such as the International Classification of Functioning, Disability and Health (ICF) that we haven’t yet investigated. We anticipate that transcribed data will impose other source of problems due to the nature of how spoken language is transformed into written form. It might be fairly cumbersome to capture patients' perceptions of health-related problems in a simple straightforward manner. A general language, near synonym dictionary should also be worth to investigate since there are numerous cases that could be captured by such resources such as *smärta* ‘pain’, *ont* ‘hurt’, *värk* ‘pain’, and enhance controlled vocabularies in order to achieve better matching. In the same spirit Zeng & Tse (2006) discuss the development of consumer health vocabularies that would reflect the different ways consumers express and think about their health is necessary for extending research on various types of information-based tools. Such resources would be also beneficial as a complement to controlled vocabularies, and particularly for health information retrieval and understanding applications. The results should serve as a useful model, both for distributed input to the enhancement of controlled vocabularies and for devising new and better means for achieving better coverage.

Acknowledgments

This work is supported by the Gothenburg Centre for Person-Centred Care (GPCC) & the CLT.

References

Brown PJ and Odusanya L. 2001. Does size matter?--Evaluation of value added content of two decades of successive coding schemes in secondary care. *Proc AMIA Symp*. 71–75.

Grishman R., Huttunen S. and Yangarber R. 2002. Information Extraction for Enhanced Access to Disease Outbreak Reports. *J Biomed Inf – Special issue: Sublanguage*. Vol. 35(4): 236-246.

Kohler M. 2008. *Unified Medical Language System for Information Extraction*. VDM Verlag.

Kokkinakis D. 2009. Lexical granularity for automatic indexing and means to achieve it – the case of Swedish MeSH®. In *Information Retrieval in Biomedicine: NLP for Knowledge Integration*. Prince V. & Roche M. (eds). Pp:11-37. IGI Global.

Krauthammer M. and Nenadic G. 2004. Term identification in the biomedical literature. *J Biomed Inf*. 37(6):512-26.

KSH97/ICD-10. 2002. Klassifikation av sjukdomar & hälsoproblem 1997; Rev 2002/04. Socialstyrelsen. <http://www.socialstyrelsen.se/Lists/Artikelkatalog/Attachments/10871/2002-4-2_200242.pdf> .

Langlotz CP. and Caldwell SA. 2005. Completeness of existing lexicons for representing radiology report information. *J Digit Imag.* 15 Suppl 1:201-5.

Tsujii J. and S. Ananiadou. 2005. Thesaurus or Logical Ontology, Which One Do We Need for Text Mining? *J. Lang Res & Eval*. Pp:77-90. Vol. 39:1.

UMLS <www.nlm.nih.gov/pubs/factsheets/umls.htm>

Wasserman H. and Wang J. 2003. An Applied Evaluation of SNOMED CT as a Clinical Vocabulary for the Computerized Diagnosis and Problem List. *AMIA Annu Symp Proc*. 699–703.

Zeng QT. and Tse T. 2006. Exploring and Developing Consumer Health Vocabularies. *J Am Med Inform Assoc.* 13:24-29.

Appendix A1

< http://www.1177.se//Fakta-och-rad/Sjukdomar>
Astma </Astma/>
Blindtarmsinflammation </Blindtarmsinflammation/>
Blodpropp i benet </Blodpropp-i-benet/>
Bältros </Baltros/>
Gallsten </Gallsten/>
Havandeskapsförgiftning </Havandeskapsforgiftning/>
Hjärtsvikt </Hjartsvikt/>
Klamydia </Klamydia/>
Laktosintolerans </Laktosintolerans/>
Ménières sjukdom </Menieres-sjukdom/>
Näthinneavlossning </Nathinneavlossning/>
Påssjuka </Passjuka/>
Rabies </Rabies/>
Ulcerös kolit </Ulceros-kolit/>
Urinvägsinfektion </Urinvagsinfektion/>
<http://www.praktiskmedicin.com/>
Akut lymfatisk leukemi <sjukdom.asp?sjukdid=897>
Analfissurer <sjukdom.asp?sjukdid=309>
Bronkit. Luftrörskatarr < sjukdom.asp?sjukdid=10>
Demens < sjukdom.asp?sjukdid=90>
Diabetes ketoacidios < sjukdom.asp?sjukdid=744>
Järnbristanemi < sjukdom.asp?sjukdid=900>
Kol. Emfysem. Kroniskt Obstruktiv Lungsjukdom < sjukdom.asp?sjukdid=14>
Lungödem < sjukdom.asp?sjukdid=147>
Njursten < sjukdom.asp?sjukdid=469>
Osteoporosis < sjukdom.asp?sjukdid=98>
Polyneuropati < sjukdom.asp?sjukdid=369>
Prostatacancer < sjukdom.asp?sjukdid=670>
Psoriasis < sjukdom.asp?sjukdid=234>
Soleksem < sjukdom.asp?sjukdid=239>
TBE-infektion < sjukdom.asp?sjukdid=1158>
<http ://www.netdoktor.se/>
ADHD <adhd/?_PageId=113320>
Artros <artros/?_PageId=162>
Bihåleinflammation <forkylning-infektion/?_PageId=505>
Cushings syndrom <hud-har/?_PageId=524>
Diskbråck <smarta/?_PageId=360>
Enterohemorragisk E. Coli (EHEC) <mage-tarm/?_PageId=550>
Fönstertittarsjuka (claudicatio intermittens) <hjart-karl/?_PageId=107115>
Genital Herpes <sex-relationer/?_PageId=432>
Hemorrojder <mage-tarm/?_PageId=583>
Irriterad tjocktarm (Colon Irritabile/IBS) <mage-tarm/?_PageId=509>
Kolera <mage-tarm/?_PageId=622>
Multipel skleros (MS) <neurologi/?_PageId=652>
RS-virus <barn/?_PageId=713>
Skrumplever (levercirrhos) <mage-tarm/?_PageId=630>
Vinterkräksjukan <mage-tarm/?_PageId=694>

Appendix A2

Reference list (top occurrences)
8 feber
6 diarré
5 trötthet
5 kräkningar
4 ångest
3 trött
3 sveda
3 smärta
3 magsmärtor
3 förvirring
2 ökad törst
2 yrsel
2 vätskeförlusten
2 viktminskning
2 tryck på ryggmärgen
2 tinnitus
2 smärtor
2 oro
2 ont i magen
2 nedsatt vibrationssinne
2 muskelsvaghet
2 medvetandesänkning
2 lätt feber
2 kramper
2 koncentrationssvårigheter
2 kallsvett
2 impotens
2 hög feber
2 hematuri
2 gaser i magen
2 förstoppning
2 dålig aptit
2 dyspné
2 depression
2 blåskatarr
2 blekhet
2 benskörhet
1 övergående ospecifik feber
1 överaktivitet
1 ömt över gallblåsan
1 ömma öronspottkörtlar
1 ömhet runt naveln
1 ökad trötthet
1 ökad hårväxt
1 ögonvitan blir gul
1 ögat kännas torrt
1 ögat bli rött
1 ögat blir känsligt för ljus
1 ödem
1 ängslan
1 återkommande trötthet
1 åldrandet
1 åderbråck i matstrupen
1 ytsensibilitet
1 vätska samlas i kroppen
…