Clinical Vocabulary and Clinical Finding Concepts in Medical Literature

Takashi Okumura
National Institute of Public Health
taka@niph.go.jp

Eiji Aramaki
Kyoto University
eiji.aramaki@gmail.com

Yuka Tateisi
National Institute of Informatics
yucca@nii.ac.jp

Abstract

Clinical decision support systems necessitate a disease knowledge base, which comprises a set of clinical findings for each disease. To efficiently represent the findings, this paper explores the relationship between clinical vocabulary and findings in medical literature through quantitative and qualitative analysis of representative disease databases. Although the data volume and the analyzed features are limited, the observations suggested the following. First, there are sets of clinical findings that are essential for physicians, but the majority of findings in medical literature are not the essential ones. Second, deviation of term frequency for clinical findings vocabulary is minimal, and clinical findings require appropriate grammar for efficient representation of findings. Third, appropriate mapping of clinical findings with clinical vocabulary would allow the efficient expression of clinical findings.

1 Introduction

Clinical decision support systems necessitate a knowledge base of diseases, which comprises efficient representations of signs and symptoms for certain diseases. Such a knowledge base may efficiently represent a disease with relation to a set of predefined findings, such as *headache* and *nausea*. However, it is commonly observed that the derivatives of such findings become a diagnostic clue in the actual diagnosis process. For example, *morning headache* may suggest a tumor in the cranium, whereas cerebral hemorrhage may accompany *sudden headache*. These cases illustrate that, in clinical medicine, signs and symptoms modified with other elements may form a meaningful cluster that carries clinically valuable information.

In order to represent the "concepts of clinical findings" in an efficient manner, we are required to maintain appropriate vocabulary, as well as a variety of modifiers, such as *where*, *when*, and *how* the signs appear. For an ontology of diseases, the analysis of the relationship between such vocabulary for clinical findings and concepts of clinical findings is indispensable for efficient knowledge representation.

Accordingly, the paper performs quantitative and qualitative analysis of the vocabulary and the concepts of clinical findings in a couple of representative disease databases, OMIM (Online Mendelian Inheritance in Man) (McKusick, 2007) and Orphanet (Aymé and Schmidtke, 2007). In Section 2, we analyze the vocabulary of clinical findings, by assessing the impact of the vocabulary size against the coverage of words in descriptions of diseases. In Section 3, variations of clinical findings concepts are analyzed by assessing the expressions of clinical findings in the same texts. These analyses are followed by Section 4, which discusses the observations of the preceding sections. Section 5 summarizes a survey of related work, and Section 6 concludes the paper.

2 Variation of clinical findings vocabulary

In this section, the distribution of the terms for clinical findings is measured by taking simple statistics of terms used in OMIM and Orphanet. OMIM contains descriptions of approximately two thousand diseases in free format texts, and Orphanet has six thousand entries for diseases, including rare diseases. In the processing, MetaMap (Aronson and Lang, 2010; Aronson, 2001) is first applied to the texts, to extract the terms related to clinical findings. MetaMap is a tool to map phrases in a given medical literature text with UMLS (Unified Medical Language System) terminology (Lindberg et al., 1993), coupled with

International Joint Conference on Natural Language Processing Workshop on Natural Language Processing for Medical and Healthcare Fields, pages 7–13,
Nagoya, Japan, 14-18 October 2013.

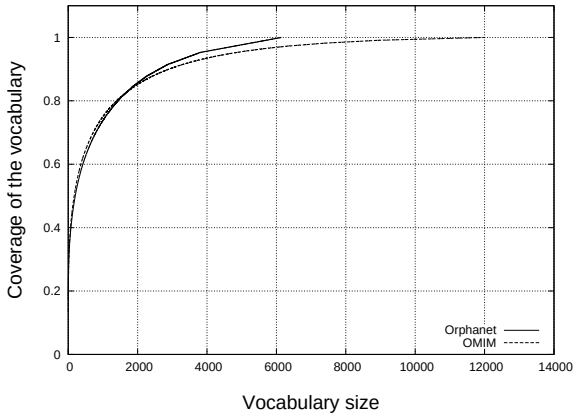

Figure 1: Vocabulary size and word coverage

Category	#	Ratio	Cumulative
Noun	254	18.6%	18.6%
Phrase			
Concept	476	34.8%	53.4%
Set of concepts	583	42.6%	96.0%
Sentence	55	4.0%	100.0%
Total	1368	100.0%	

Table 1: Grammatical categories of clinical findings in OMIM 20 documents

their semantic category. Then, the following subject categories of UMLS terms are used to extract clinical findings in the text: Acquired Abnormality, Anatomical Abnormality, Cell or Molecular Dysfunction, Congenital Abnormality, Disease or Syndrome, Embryonic Structure, Injury or Poisoning, Mental or Behavioral Dysfunction, Neoplastic Process, Sign or Symptom, and Virus. Lastly, the frequency of the extracted terms is measured, and the coverage of the symptomatic terms in the documents is calculated by increasing the vocabulary size in order of the term frequency.

Figure 1 shows the cumulative distribution of the terms. As illustrated, the coverage of the terms increases by adding terms for clinical findings into the vocabulary, and the top 2000 words covers 85% of the terms for clinical findings in the databases. Beyond this point, the coverage becomes less responsive to the addition of terms, because they are infrequently used in the target documents. The figure suggests that the difference between OMIM and Orphanet for the observed tendency is minimal.

To assess the size constraint of the vocabulary, we measured the percentage of simple words in the description of diseases. A clinical finding can be a word, such as *fever*, a phrase, such as *periodic fever*, or a sentence. If the portion of word findings is limited among all expression types, the unlimited vocabulary size by itself cannot achieve the appropriate representation of clinical findings. To estimate the upper bound of the contribution by the unlimited vocabulary, we analyzed the findings described in randomly selected OMIM documents (Document IDs: 108450, 113450 118450, 123450, 140450, 176450, 181450, 200450, 203450, 214450, 218450, 233450 236450, 244450, 248450, 259450, 265450, 267450, 305450, and 311450). The 20 texts included 1368 clinical findings in total. A sample phrase and a sentence are excerpted below:

"with most patients dying within 10 years of onset"
(OMIM 203450: Alexander Disease)

"No females manifested any symptoms."
(OMIM 305450: Opitz-Kaveggia Syndrome)

Table 1 shows the breakdown of the finding categories in the selected texts. The "Noun" category is for single word. The "Phrase" category is for phrases, which comprise phrases that represent either a concept, or a set of concepts. "Concept" includes phrases that can be mapped to a clinical concept, such as "mental retardation" and "Tetralogy of Fallot". "Set of concepts" is for phrases that are mapped to multiple concepts. As the table illustrates, the noun category accounts for only 18.6% of the expressions for clinical findings. Even if appropriate terminologies cover simple concepts for clinical findings (34.8%), they share only 53.4% of the entire findings and 46.6% of the expressions still necessitate phrases and sentences. Although the distinction of a concept and a set of concepts can be ambiguous in some cases, this tendency suggests that even the unlimited vocabulary cannot appropriately express all the clinical findings because the portion for vocabulary is limited in the actual descriptions of diseases.

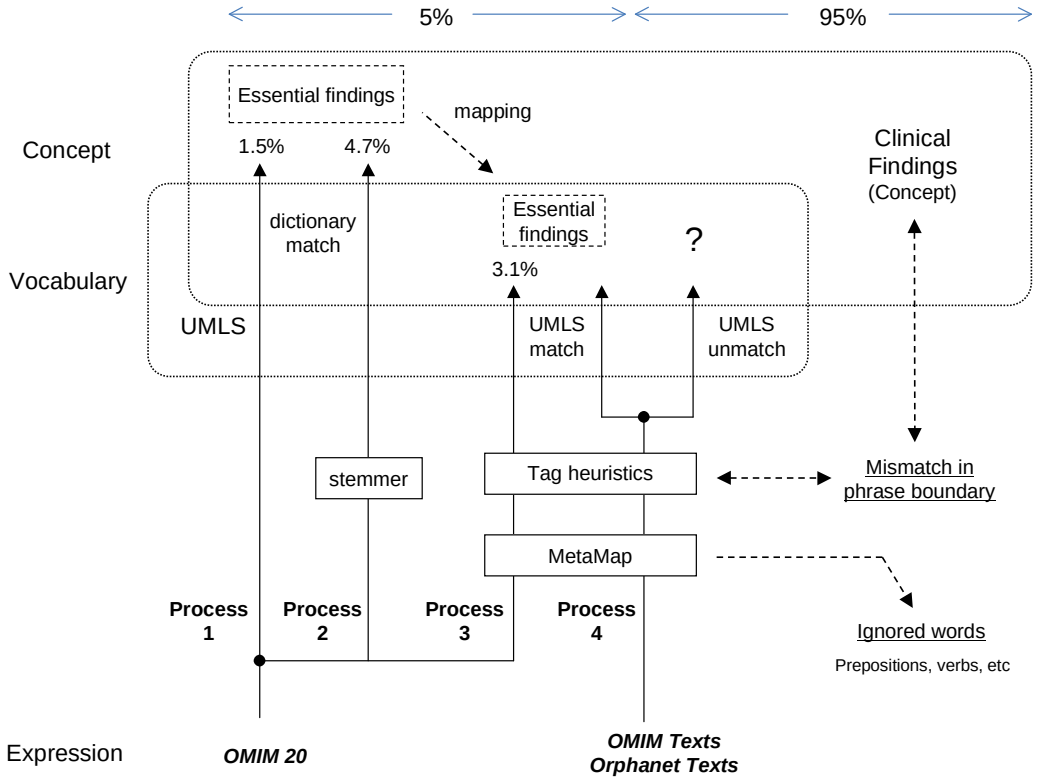

Figure 2: Coding of clinical findings concepts and the vocabulary

3 Variation of clinical findings concepts

The last section explored the vocabulary, required to express clinical findings. This section, inversely, examines clinical finding concepts in the descriptions of diseases. For this purpose, we used a set of clinical findings, compiled by Dr. Keijiro Torigoe for his rudimentary diagnostic reminder system (Torigoe et al., 2003). The dataset comprises 597 clinical manifestations, which include common signs and symptoms as well as entries for typical laboratory examinations results, such as high white blood cell counts and low platelets. The analysis in this section utilizes the essential findings for physicians to code clinical findings in the annotated OMIM texts. The entire setting is illustrated in Figure 2.

First, we performed dictionary matching of the essential finding, against the annotated OMIM texts (Process 1, Figure 2). The simple dictionary match showed that the essential findings accounted for only 1.5% of the annotated elements. Second, we applied a stemmer, Snowball (Porter, 2001), before the matching, which increased the recall to 4.7% (Process 2, Figure 2). Third, for further performance gain, we processed the essential findings with MetaMap and compiled the set of 586 findings in *UMLS Concept Unique Identifiers (CUI)*. Then, we performed the matching against the result of the MetaMap processing on the OMIM texts, which resulted in a 3.1% match (Process 3, Figure 2).

The three stages illustrated that the essential findings account for only 5% of the clinical findings in the sample documents, and the failure analysis suggested the following. First of all, MetaMap mostly ignores prepositions and verbs, which constitute essential parts in the expression of the clinical findings. Second, MetaMap segments the texts into minimum phrases that have corresponding CUI. However, annotators with a clinical background tend to group multiple phrases together, because they carry meaningful information as clinical findings. This results in the further mismatch between the MetaMap output and the concepts of clinical findings. Lastly, the dataset of essential findings could have missed some frequent terms.

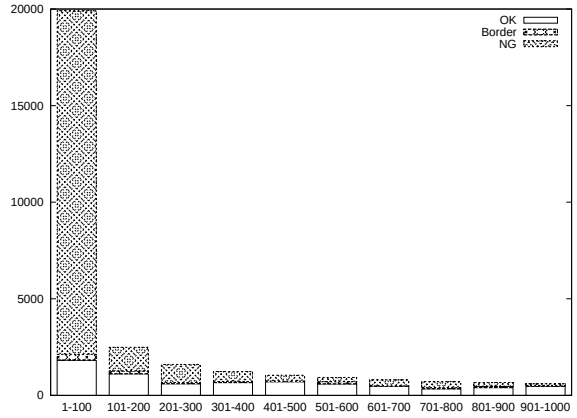

Figure 3: Word count of unmatched terms in Orphanet

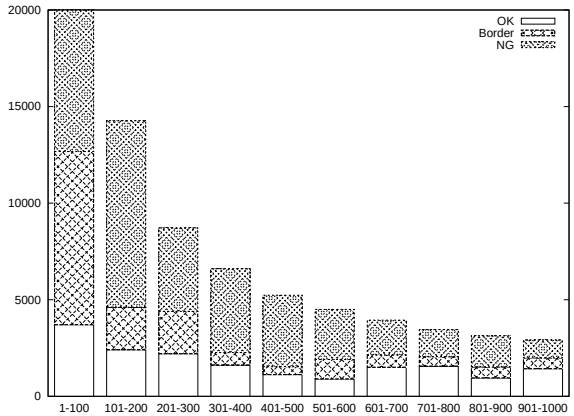

Figure 4: Word count of unmatched terms in OMIM

To verify the last conjecture and to detect unknown frequent findings, we analyzed the failure cases of UMLS matching. For this purpose, we applied the same processing on the entire OMIM and Orphanet texts and extracted unmatched cases (Process 4, Figure 2). The output was sorted in order of frequency and grouped into 10 clusters, each of which contained 100 words. Figures 3 and 4 show the word count of terms in each cluster (Note the range of y-axis is limited). As illustrated, the first clusters of both graphs exhibited striking peaks (Orphanet: 19898 counts and OMIM: 86085 counts) for the top 100 words. However, a detailed look revealed that half of the terms were useless (NG class), because they are functional terms that were mistakenly included by the tag heuristics. The class Border denotes borderline cases, which are clinical terms, but which do not carry clinical meaning in the context, such as *(the) disease* and *(the) symptom*.

Accordingly, we extracted the acceptable OK terms in the top 1000 unmatched words and measured their contribution to the frequency distribution of the terms in the target documents. As Figure 5 illustrates, the contribution of the unmatched words is limited: the top 100 unmatched words accounted for 6.8% of the findings word count for Orphanet documents, and 3.1% of the findings word count for OMIM documents. In all, the 1000 unmatched words contained 543 words for Orphanet and 278 words for OMIM, which accounted for 14.6% and 4.8% in the entire word count, respectively. The Orphanet cases outperformed the OMIM cases, which could be partially attributed to the difference in word counts (49,342

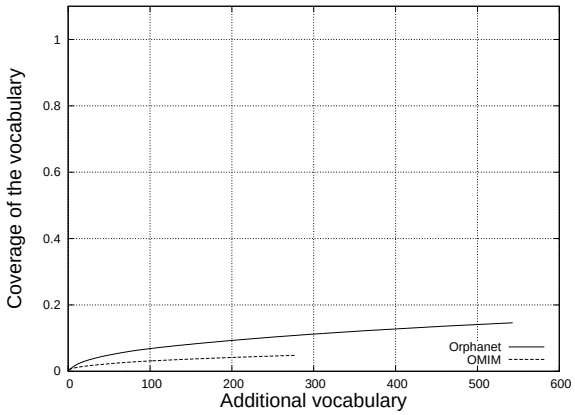

Figure 5: Additional vocabulary and contribution to the word coverage

against 361,566). The majority of the cases in the frequent unmatched words were technical terms, such as *microcephaly*, *hypertelorism*, and *facial dysmorphism*, which could be frequent for certain classes of genetic disorders but clinically uncommon.

The observation suggests that the number of essential clinical findings is approximately several hundreds, far below a thousand. These findings account for just a few percent of clinical findings documented in the descriptions of diseases in representative disease databases. The concepts of clinical findings in medical literature are diverse. Although some of the clinical finding concepts might be well-known, such as anatomical and congenital anomalies, most of them are clinically uncommon and do not appear often in literature. Although manual matching might increase the percentage, it is not likely that the overall picture would change.

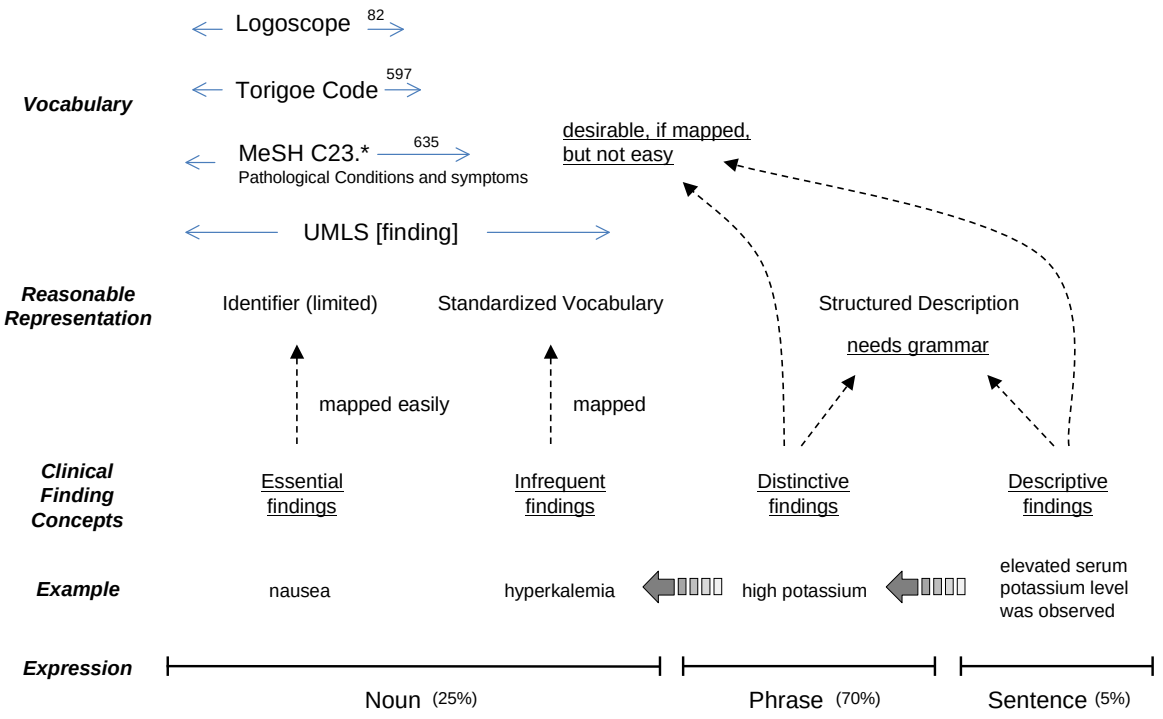

Figure 6: Clinical findings concepts and vocabularies

4 Discussion

This paper explores the relationship between clinical vocabulary and clinical findings through the analysis of disease descriptions compiled in OMIM and Orphanet. Figure 6 summarizes the observations. The essential findings, found as nouns in literature, are limited in number. This is supported by other datasets. For example, Logoscope (Nash, 1960) is a manually operated diagnostic tool called the "diagnostic slide rule", and it defines a set of 82 essential findings. MeSH (National Library of Medicine, 1963) has a category for essential findings named "pathological conditions and symptoms", which includes 635 terms. Infrequent findings are the other type of nouns, which can be mapped to corresponding terms in standardized vocabulary. There are other types of clinical findings, distinctive and descriptive findings, although the classification is tentative. Distinctive findings are those expressed by phrases, and descriptive findings are the most complex findings, which can be expressed only by sentences.

It is desirable that the findings are properly mapped onto the clinical findings vocabulary. However, it is not an easy task to map them, as suggested throughout the paper. For appropriate representation of clinical findings concepts, vocabulary alone cannot bridge the gap, and a grammar with appropriate descriptive power is indispensable. However, there is a tradeoff between the descriptive power and the cost for knowledge acquisition and representation.

mental retardation
intellectual deficiency
intellectual impairment
language delay
learning difficulties
poor school performance
delayed psychomotor development
psychomotor delay
psychomotor retardation

Table 2: Expressions for mental retardation

The knowledge acquisition of clinical findings is still a challenging task for Natural Language

Processing (NLP). Physicians may describe a finding in a sentence, which is common for pathological and radiological findings. Such a finding might have a corresponding term, and an elaborated system might cleverly map the sentence into a standardized vocabulary. However, this process involves various tasks, such as processing of negation, dependency, ambiguity, and abstraction, most of which are still unreliable for clinical use at this point. Even mapping of phrases is a challenge. For example, physicians may describe the concept "mental retardation" in diverse ways. Table 2 denotes how the concept is expressed in OMIM and Orphanet. Although MetaMap is a useful tool for mapping clinical terms, it still falls short of the required performance, to map sentences and phrases into standardized vocabulary.

The high cost of knowledge acquisition also applies to knowledge representation. Structured description of knowledge requires a grammar, which also burdens the data utilization process. Accordingly, it would be beneficial to reduce the cost for data representation, in addition to the improvement of knowledge acquisition performance. In this regard, physicians may express findings in phrases and sentences, when the findings are unfamiliar, or when they do not recall the appropriate terms even if one exists that corresponds to the concept. Examples include laboratory findings as illustrated at the bottom of Figure 6. Because a noun is the simplest form of knowledge, mapping of sentences and phrases into the terms might contribute to reducing the representation cost as well as the cost for data utilization.

5 Related works

Numerous research efforts have been made in the field of Natural Language Processing toward precise acquisition and representation of knowledge in clinical medicine.

First of all, there is a class of works aimed at boosting the accuracy of finding clinical manifestations in medical texts. Since the pioneering work (Sneiderman et al., 1996) in this problem domain, there have been various studies that investigated basic technologies required for accurate mining. Chapman proposed negation detection (Chapman et al., 2001) for clinical texts, which was extended to context handling methods (Chapman et al., 2007). Other groups focused on knowledge acquisition of diseases (Achour et al., 2001; Aleksovska-Stojkovska and Loskovska, 2010).

Second, researchers have worked on the knowledge representation issue for clinical findings. In addition to UMLS (Lindberg et al., 1993), which is used in this paper, SNOMED-CT (International Health Terminology Standard Development Organisation, 2001), MeSH (National Library of Medicine, 1963), OpenGALEN (Rector et al., 2003), and MedDRA (MedDRA Maintenance and Support Services Organization, 2007) have been used for the representation of clinical concepts. There are other studies in this domain (Sager et al., 1994; Cimino, 1991; Kong et al., 2008; Peleg and Tu, 2006).

Lastly, there have been several lines of work that explored the tools for information extraction on clinical reports. For example, (Friedman et al., 1994) developed Medical Language Extraction and Encoding (MedLEE) to encode clinical documents in a structured form. The Mayo Clinic also developed a similar NLP system (cTakes) (Savova et al., 2010) for clinical reports and TEXT2TABLE (Aramaki et al., 2009) targeted Japanese discharge summaries.

6 Conclusion

This paper investigated clinical vocabulary and clinical finding concepts. Because the analysis is made with a limited set of data, further study is required for more rigorous proof. Nevertheless, the current observations suggest the following.

First, there are essential findings for physicians and, in medical literature, the majority of the findings do not fall into the category. This observation is consistent with the fact that annotated findings tend to span multiple words.

Second, the deviation of the term frequency for clinical findings vocabulary is minimal, and the vocabulary alone cannot express all the clinical findings. Even with the UMLS terminology, the expressive power is limited, which necessitates an appropriate grammar for structured descriptions of findings.

Third, knowledge acquisition of clinical findings is costly, and the grammar would escalate the cost for representation, as well as the cost for data utilization. However, appropriate mapping of clinical findings and clinical vocabulary, particularly for infrequent terms, might contribute toward expressing clinical findings without increasing the cost for representation and for utilization.

References

Soumeya L Achour, Michel Dojat, Claire Rieux, Philippe Bierling, and Eric Lepage. 2001. A umls-based knowledge acquisition tool for rule-based clinical decision support system development. *Journal of the American Medical Informatics Association*, 8(4):351–360.

Liljana Aleksovska-Stojkovska and Suzana Loskovska. 2010. Clinical decision support systems: Medical knowledge acquisition and representation methods. In *2010 IEEE International Conference on Electro/Information Technology (EIT)*, pages 1–6. IEEE.

Eiji Aramaki, Yasuhide Miura, Masatsugu Tonoike, Tomoko Ohkuma, Hiroshi Mashuichi, and Kazuhiko Ohe. 2009. Text2table: Medical text summarization system based on named entity recognition and modality identification. In *the Human Language Technology conference and the North American chapter of the Association for Computational Linguistics (HLT-NAACL2009) Workshop on BioNLP*, pages 185–192.

Alan R Aronson and Francois-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–36.

Alan R Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *AMIA Annual Symposium*, pages 17–21.

Ségolène Aymé and Jorg Schmidtke. 2007. Networking for rare diseases: a necessity for europe. *Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz*, 50(12):1477–1483.

Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310.

Wendy W Chapman, John Dowling, and David Chu. 2007. Context: An algorithm for identifying contextual features from clinical text. In *Biological, translational, and clinical language processing (BioNLP2007)*, pages 81–88.

James J Cimino. 1991. Representation of clinical laboratory terminology in the unified medical language system. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 199. American Medical Informatics Association.

Carol Friedman, Philip O Alderson, John HM Austin, James J Cimino, and Stephen B Johnson. 1994. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, 1(2):161–174.

International Health Terminology Standard Development Organisation. 2001. SNOMED CT. http://snomed.org/.

Guilan Kong, Dong-Ling Xu, and Jian-Bo Yang. 2008. Clinical decision support systems: a review on knowledge representation and inference under uncertainties. *International Journal of Computational Intelligence Systems*, 1(2):159–167.

Donald Lindberg, Betsy Humphreys, and Alexa McCray. 1993. The unified medical language system. *Methods of Information in Medicine*, 32(4):281–91.

Victor A. McKusick. 2007. Mendelian inheritance in man and its online version, omim. *American journal of human genetics*, 80(4):588–604, April.

MedDRA Maintenance and Support Services Organization. 2007. *Introductory Guide, MedDRA Version 10.1*. International Federation of Pharmaceutical Manufacturers and Associations.

Firmin A. Nash. 1960. Diagnostic reasoning and the logoscope. *Lancet*, 276:1442–1446, December.

National Library of Medicine. 1963. Medical Subject Headings. http://www.nlm.nih.gov/mesh/.

Mor Peleg and Samson Tu. 2006. Decision support, knowledge representation and management in medicine. *Yearb Med Inform*, 45:72–80.

Martin Porter. 2001. Snowball: A language for stemming algorithms.

Alan Rector, Jeremy Rogers, Pieter Zanstra, and Egbert van der Haring. 2003. Opengalen: open source medical terminology and tools. In *AMIA Annual Symposium Proceedings*, page 982.

Naomi Sager, Margaret Lyman, Christine Bucknall, Ngo Nhan, and Leo J Tick. 1994. Natural language processing and the representation of clinical data. *Journal of the American Medical Informatics Association*, 1(2):142–160.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Charles A Sneiderman, Thomas C Rindflesch, and Alan R Aronson. 1996. Finding the findings: identification of findings in medical literature using restricted natural language processing. In *AMIA Annual Fall Symposium*, pages 239–43. AMIA.

Keijirou Torigoe, Gen'ichi Kato, and Yoshio Ohta. 2003. Computer-aided diagnosis supporting tool. *Japan Medical Journal*, (4120):24–32. (in Japanese).