# Using UMLS Semantics for Classification Purposes

**Olivier Bodenreider, M.D., Ph.D.**
**National Library of Medicine, Bethesda, Maryland**
olivier@nlm.nih.gov

*The Unified Medical Language System (UMLS) contains semantic information about terms from various sources; each concept can be understood and located by its relationships to other concepts. We describe a method in which the semantic relationships between UMLS concepts are exploited for the purpose of classification. This method combines three existing components: 1) Mapping terms to UMLS concepts; 2) Restricting UMLS concepts to MeSH; and 3) Mapping MeSH terms to disease categories. When applied to the automatic classification of condition terms into broad disease categories in the Clinical Trials database, this method assigned relevant categories to 92% of the 1823 condition terms encountered. 135 (7%) failed to be classified and 14 (.77%) were misclassified. The limits of this method are discussed, as well as the reuse of existing components, and the tuning required to achieve automatic classification.*

## INTRODUCTION

In patient- or consumer-oriented health information systems, such as MEDLINE*plus*[1] or the Clinical Trials database[2], condition terms are indexed by broad disease categories such as "Eye Diseases" or "Parasitic Diseases," allowing users to navigate the system in browse mode. A condition term may be assigned to several disease categories, increasing the possibility of retrieving a given condition from different categories. For example, the term "adrenal medulla neoplasm" (a tumor of adrenal gland) could be assigned to both "Endocrine Diseases" and "Neoplasms" categories. Dynamic systems in which data may be added on a continuing basis require condition terms to be classified automatically, with no misclassified conditions and few non-classified conditions. Our goal is to develop a method whereby specific disease names (or, more generally, names for medical conditions), referred to as condition terms, can be automatically classified into broad disease categories.

Traditional classification methods such as statistical techniques or neural networks may be used for such a task. These methods rely on modeling the association between condition terms and disease categories from a training set. As an alternative, we decided to exploit the semantic properties of the Unified Medical Language System® (UMLS®) and to explore the possibility of using inter-concept relationships in the UMLS to select disease categories found in the semantic vicinity of a given condition term. This approach has already been used successfully in the Indexing Initiative (IND), an ongoing effort of the National Library of Medicine to investigate automated indexing methods as a partial or complete substitute for current indexing practices [1]. In the IND project, nominal phrases are extracted from medical text and mapped to UMLS concepts; concepts are then restricted to the Medical Subject Headings® (MeSH) vocabulary. Finally, MeSH descriptor candidates are ranked for how well they represent the content of the input text.

Compared to IND, the classification of condition terms in disease categories is expected to be a somewhat easier task:

- A list of condition terms, even when unrestrained, represents a relatively limited set of concepts, whereas noun phrases extracted from arbitrary text may be more diverse;
- Condition terms are generally well represented in the UMLS, coming for example from clinically-oriented vocabularies such as SNOMED or Clinical Terms Version 3 (Read Codes);
- Finally, it is easier to map to relevant high-level categories than to find relevant descriptors most closely associated with a given concept.

These favorable conditions are expected to compensate for an additional constraint: the need for automatic classification.

After presenting the principles for using UMLS semantics to classify condition terms into broad disease categories, we will describe how this method was applied specifically to the task of classifying the condition terms found in the Clinical Trials database. Finally, we discuss the results of evaluating the methodology proposed.

## BACKGROUND

An algorithm relying on UMLS semantics to classify condition terms into MeSH disease categories is based on the following assumptions:

- It is possible to map most of the conditions terms to the UMLS;
- The several levels of organization provided by the UMLS allow for the mapping of arbitrary concepts

[1] medlineplus.gov (accessed July 17, 2000)
[2] ClinicalTrials.gov (accessed July 17, 2000)

Proceedings of AMIA Annual Symposium 2000;:86-90.

to one or more terms from the MeSH vocabulary from which disease categories are drawn;
- Specific disease names in MeSH are mapped accurately to broad disease categories.

The UMLS is intended to help health professionals and researchers use biomedical information from more than 40 vocabularies [2], including some major clinical terminologies such as SNOMED, ICD and the Read Codes, as well as MeSH, our target vocabulary. While the structure of each source vocabulary is preserved, terms that are equivalent in meaning are clustered into a unique concept. Furthermore, inter-concept relationships, either inherited from the source vocabularies or specifically generated, give the UMLS Metathesaurus additional semantic structure [3]. This structure can be visualized as a graph in which concepts are the nodes and inter-concept relationships are the links between nodes. The consequence of this is that the association between a condition term and a disease category (both nodes of the graph) is represented as a path between the two nodes in the graph, using the appropriate semantic relationships. Furthermore, the polyhierarchical structure of the MeSH vocabulary alone provides an easy mapping of any MeSH disease term to its corresponding parent categories.

| **Disease category** | **MeSH** |
|---|---|
| Bacterial and Fungal Diseases | C01 |
| Blood and Lymph Conditions | C15 |
| Cancers and other Neoplasms | C04 |
| Conditions of the Urinary Tract and Sexual Organs, and Pregnancy | C12, C13 |
| Digestive System Diseases | C06 |
| Diseases and Abnormalities at or before Birth | C16 |
| Ear, Nose and Throat Diseases | C09 |
| Eye Diseases | C11 |
| Gland and Hormone Related Diseases | C19 |
| Heart and Blood Vessel Diseases | C14 |
| Immune System Diseases | C20 |
| Injuries, Poisonings, and Occupational Diseases | C21 |
| Mental Disorders | F03 |
| Mouth and Tooth Diseases | C07 |
| Muscle, Bone and Cartilage Diseases | C05 |
| Nervous System Diseases | C10 |
| Nutritional and Metabolic Diseases | C18 |
| Parasitic Diseases | C03 |
| Respiratory Tract (Lung and Bronchial) Diseases | C08 |
| Skin and Connective Tissue Diseases | C17 |
| Symptoms and General Pathology | C23 |
| Viral Diseases | C02 |

*Table 1 – List of disease categories.*

Broad disease categories such as those listed in MeSH under the term “Diseases” are suitable for our classification scheme. Mental disorders, classified elsewhere in MeSH are also of interest here. The list of disease categories used in the Clinical Trials system is given in Table 1, along with the corresponding MeSH categories.

## METHODS

The following three steps are used to classify condition terms into MeSH disease categories: condition terms are first mapped to UMLS concepts; then these concepts are restricted to the MeSH vocabulary; and finally, MeSH terms are mapped to disease categories. These methods can be combined into a strategy that maximizes the chances of finding relevant categories for condition terms.

### Mapping condition terms to the UMLS

The UMLS not only contains a large number of disease terms, including numerous synonyms and variants, but also provides lexical resources for processing medical terms. Thus, if a condition term can not be found in the UMLS through an exact match, normalization techniques (including case, punctuation, inflection and word order insensitivity) can be used to map it to the UMLS [4]. For example, the term “chromosome 4 short arm deletion” does not exist in the UMLS, but is mapped to the term “deletion of the short arm of chromosome 4” after normalization.

Among the terms that fail to map to the UMLS after normalization, some are more specific than equivalent terms in the UMLS. Removing the indicators of this specificity often makes it possible to map the input term to a concept that is broader in meaning in the UMLS, with no undesirable effects on the final classification process. For example, the term “chronic neutropenia” fails to map to the UMLS, whereas “neutropenia”, with no qualifier, is an exact match.

### Restricting UMLS concepts to MeSH

In order to restrict arbitrary UMLS concepts to the MeSH vocabulary, we reuse an algorithm that was designed to find the MeSH terms most closely associated with a UMLS concept for the purpose of automatic indexing of medical texts [5]. This algorithm exploits several UMLS semantic properties, including synonymy, inter-concept relationships and the categorization of concepts. The overall strategy involves the following four steps:

1. Choose a MeSH term as a synonym of the initial concept.
2. Choose an associated expression which is the translation of the initial concept.
3. Select MeSH terms from concepts hierarchically related to the initial concept
4. Base the selection on the non-hierarchically related concepts of the initial concept.

The algorithm stops at any step that succeeds.
For example, the condition term "cancrum oris" is directly mapped to the MeSH term "Noma", both being synonyms in the UMLS. The condition term "neurogenic hypertension", a UMLS concept, is mapped to the MeSH term "Hypertension" which is hierarchically related to it in the UMLS.

**Mapping MeSH descriptors to disease categories**

Once mapped to a MeSH descriptor, the polyhierarchical structure of MeSH can be exploited to assign broad categories to the original term. In MeSH, each descriptor has both a unique identifier and one or more tree numbers used to describe the hierarchical structure of the vocabulary. The tree numbers reflect the nodes between the root and a given term. These allow particular trees such as the "Diseases" tree (tree numbers starting by "C") to be easily identified. In addition, mental disorders, classified in MeSH under the F03 tree, are also a disease category in our system.
For example, the MeSH term "Adrenal Gland Neoplasms" has unique identifier "D000310" and tree numbers "C04.588.322.78", "C19.53.347" and "C19.344.78". The disease categories relevant to this term are "Neoplasms" (C04) and "Endocrine Diseases" (C19); they can be computed from the tree number by extracting the left-most node.

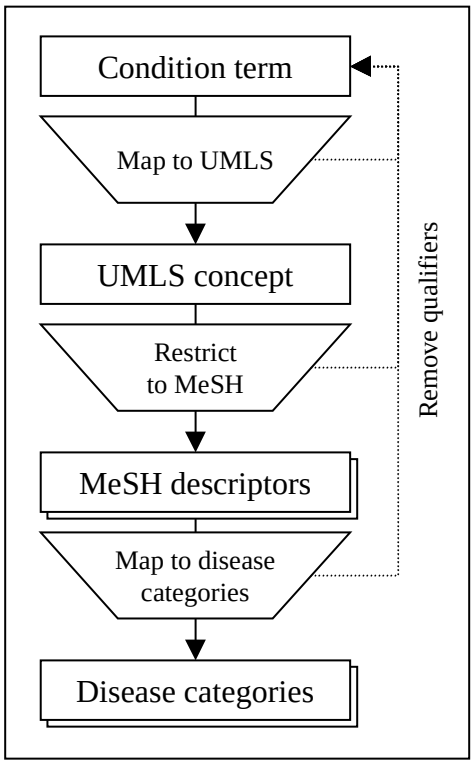


*Figure 1 – Classification strategy.*

**Classification strategy**

The strategy for classifying condition terms into disease categories involves the following three steps (presented earlier), as shown in Figure 1:

1. **Map to the UMLS**. Three progressive levels of aggressiveness are applied to the condition term: exact match, normalization, removal of qualifiers.
2. **Restrict to MeSH**. If the UMLS concept is not restricted to MeSH, the process is started again from the beginning after qualifiers have been removed from the condition term.
3. **Map to disease categories**. If none of the MeSH descriptors belong to one of the relevant trees, the process is started again from the beginning after qualifiers have been removed from the condition term.

For example, the condition term "Stage II multiple myeloma" is an exact match to a UMLS concept, but fails to map to a MeSH descriptor. After removing the qualifier "Stage II", the term "multiple myeloma" correctly maps to the relevant categories ("Cancers and other Neoplasms", "Heart and Blood Vessel Diseases", "Blood and Lymph Conditions" and "Immune System Diseases").
The list of qualifiers that may to be removed from terms is given in Table 2. Any mention of "phase" or "stage" is also considered removable.

| | | |
|---|---|---|
| acquired | infant | recurrent |
| acute | juvenile-onset | risk reduction |
| adult | mild | secondary |
| age-related | newly diagnosed | severe |
| childhood | prevention of | unspecified |
| chronic | previously treated | untreated |
| congenital | primary | phase X |
| idiopathic | primitive | stage X |

*Table 2 – List of qualifiers that may be removed from condition terms.*

**Evaluation**

The classification algorithm was applied to the 12,612 condition terms (1823 different terms) used to describe the 4423 studies currently in the Clinical Trials database.
The quality of the classification process was evaluated both at each step of the process and globally. The quality of the overall classification process was evaluated by manual review.

## RESULTS

Out of the 1823 condition terms, 1746 (96%) were associated with at least one descriptor in MeSH.

Tables 3 and 4 show the details of the methods used to map condition terms to MeSH, sorted by ascending order of aggressiveness.

| Method | N | % |
|---|---|---|
| Exact Match | 1694 | 97 % |
| Normalized String Index | 14 | 1 % |
| Removal of Qualifiers | 38 | 2 % |
| Total | 1746 | 100 % |

*Table 3 – Mapping condition terms to the UMLS.*

| Method | N | % |
|---|---|---|
| Synonymy | 930 | 53 % |
| Associated Expressions | 36 | 2 % |
| Hierarchically related concepts | 780 | 45 % |
| Total | 1746 | 100 % |

*Table 4 – Restricting UMLS concepts to MeSH.*

Causes of failure to map condition terms to the UMLS include unusually qualified terms (e.g. "First-episode schizophrenia"), insufficiently qualified terms (e.g. "Type 2 Diabetes [mellitus]"), unusual eponymic terms (e.g. "Smith-Magenis syndrome"), as well as complex or unusual terms.
Causes of failure to restrict UMLS concepts to MeSH include the fact that some relationships are not represented in the UMLS (e.g. the DSM IV term "Cognitive Disorders" is unrelated to the MeSH term "Cognition Disorders").

Out of the 1746 condition terms associated with at least one descriptor in MeSH, 1688 (97%) were mapped to at least one disease category. The distribution of the number of categories mapped to is presented in Table 5.

| Number of categories | N | % |
|---|---|---|
| 1 single category | 737 | 44 % |
| 2 different categories | 501 | 30 % |
| 3 different categories | 359 | 21 % |
| 4 or more categories | 91 | 5 % |
| Total | 1688 | 100 % |

*Table 5 – Mapping MeSH descriptors to disease categories.*

| Relevance | N | % |
|---|---|---|
| Fully relevant | 1514 | 90 % |
| Partially relevant | 160 | 9 % |
| Non-relevant | 14 | 1 % |
| Total | 1688 | 100 % |

*Table 6 – Overall classification process.*

In the classification, the evaluation of relevance presented in Table 6 was performed as follows. "Fully relevant" means that neither of the following situations occur: 1) a non-relevant disease category is associated with the condition term ("Non-relevant"), or 2) a relevant category is missing ("Partially relevant").
Most cases of partially relevant classification involve cancer terms for which only the "Cancers and other Neoplasms" category is selected, whereas the category corresponding to the location of the cancer is missing. For example "Astrocytoma", a brain cancer, fails to be classified in the "Nervous System Diseases" category. This reflects how the corresponding concepts are represented in the UMLS. Terms describing various forms of leukemia account for a third of these cases. Misclassified terms (classified in a non-relevant category) are rare and are due essentially to ambiguity in the UMLS.

## DISCUSSION

### Reuse of existing components

The mapping of text to the UMLS has long been recognized as a feature needed in various applications. Resources such as the various indexes built from UMLS strings are part of the standard UMLS distribution [3]. More sophisticated but less portable programs (e.g. MetaMap [6]) are made available through the UMLS Knowledge Source Server, allowing for approximate matching.
Mapping between vocabularies is also necessary wherever different vocabularies or different versions of a given vocabulary are in use. Such a mapping is often implemented through fixed tables. The ability to map vocabularies automatically through semantic properties, beyond the presence of synonym terms, reduces the cost of having human coders develop mapping tables.
In this experiment, a major concern has been to reuse existing components (software and algorithms) instead of developing ad hoc tools. The mapping of condition terms to the UMLS uses UMLS indexes. The algorithm designed to restrict UMLS concepts to MeSH for the Indexing Initiative project proves to be useful in the context of classifying condition terms.
Adaptations of the original components and methods are presented next.

### Adaptation for automatic classification

Compared to other applications in which these components may be used, the classification of condition terms into disease categories presents several particularities:

1. Condition terms, or keywords used to describe clinical trials, constitute a sort of controlled vocabulary, more limited and closer to existing vocabularies than noun phrases extracted from

medical journal articles, for example. Clinical trials provided by the National Cancer Institute (NCI) use terms from NCI's Physician Data Query (PDQ) thesaurus, one of the vocabularies in the UMLS. Therefore, mapping condition terms to the UMLS requires less aggressive lexical techniques to achieve satisfactory performance. On the other hand, the granularity of some condition terms is greater than the granularity of corresponding UMLS terms. For this reason, our mapping strategy includes the removal of frequently used qualifiers, known to prevent terms from being mapped to the UMLS through simple techniques.

2. The automatic classification of condition terms requires that only relevant categories be selected, since no additional information is available to further refine them. For this reason, the mapping of condition terms to the UMLS does not use approximate matching techniques. Furthermore, the use of the algorithm restricting UMLS concepts to MeSH is tuned to favor high precision over high recall (only the first 3 steps mentioned in the Methods sections are used).

**Limits of semantics-based classification**

Classification algorithms based on semantics rely on a source of knowledge that is external to the data to be classified. Consequently, a lack of semantic relationships represented in the knowledge source might affect the performance of such an algorithm. Although the UMLS contains over 8 million pairs of related concepts, the lack of certain relationships has already been identified in other studies [7, 8].

The following example of the relation of cancer terms to neoplasm terms illustrates this issue. Cancers are usually defined as "malignant neoplasms" (Steadman's), making cancer a kind of neoplasm, from a knowledge perspective. Some relationship is therefore expected to be found in the UMLS between terms designating a cancer of a given anatomic site and a neoplasm of the same site, which is the most common situation (Table 7). There is no semantic relationship, however, between "eye cancer" and "eye neoplasm" in the 1999 edition of the UMLS[3].

| Anatomic Site | Cancer | Neoplasm | Relationship |
|---|---|---|---|
| Prostate | C0376358 | C0376358 | Synonymy |
| Liver | C0345904 | C0023903 | Parent/Child |
| Eye | C0279149 | C0015414 | None |

*Table 7 – Relationship of cancer to neoplasm terms in the UMLS according to different anatomic sites.*

[3] The missing parent/child relationship between "eye cancer" and "eye neoplasm" has been added in the 2000 UMLS.

## CONCLUSION

UMLS semantics has proven to be useful for the classification of condition terms into disease categories. This approach also benefited from reusing and adapting UMLS components. The performance of the classification algorithm is satisfactory, although 7% of the condition terms fail to be classified. Disease categories are high level descriptors and may come from several condition terms in a given clinical trial. So, setting the balance between precision and recall in favor of precision at the level of the condition terms greatly reduces the rate of misclassified condition terms, yet does not seem to be detrimental to recall at the level of the clinical trials.

Acknowledgements

The author would like to thank Dr. Alexa McCray and the members of the Clinical Trials team at the National Library of Medicine, especially Erik Dorfman, Nicholas Ide and Anthony Tse who helped define and refine the classification strategy.

References

1. Aronson AR, Bodenreider O, Chang HF, Humphrey SM, Mork JG, Nelson SJ, et al. The NML Indexing initiative. Proc. AMIA Fall Symposium 2000:(In press).
2. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med 1993;32(4):281-91.
3. UMLS. UMLS Knowledge Sources. 10th ed. Bethesda (MD): National Library of Medicine; 1999.
4. McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. Proc Annu Symp Comput Appl Med Care 1994:235-9.
5. Bodenreider O, Nelson SJ, Hole WT, Chang HF. Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. Proceedings of AMIA Annual Fall Symposium 1998:815-9.
6. Aronson AR. The effect of textual variation on concept based information retrieval. Proc AMIA Annu Fall Symp 1996:373-7.
7. Burgun A, Botti G, Bodenreider O, Delamarre D, Leveque JM, Lukacs B, et al. Methodology for using the UMLS as a background knowledge for the description of surgical procedures. Int J Biomed Comput 1996;43(3):189-202.
8. Cimino JJ. Auditing the Unified Medical Language System with semantic methods. J Am Med Inform Assoc 1998;5(1):41-51.