

Introduction

The open-access and collaborative (OAC) consumer health vocabulary (CHV) is developed through an on-going collaboration among researchers from a number of institutions including Brigham and Women's Hospital, Harvard Medical School, National Library of Medicine, and University of Wisconsin. The OAC development is driven by the needs of consumer health applications and utilizes both text analysis and human review of consumer utterances.

The OAC CHV is designed to complement existing knowledge in the Unified Medical Language System (UMLS). It differs from the UMLS (and most of its source vocabularies) in several aspects:

1. The OAC CHV focuses on expressions and concepts that are employed by health-related communications from or to consumers. As a result, it includes ambiguous, vague, slang and misspelled terms.
2. As we continue to improve the domain coverage, high-frequency terms and concepts are given priority during the development.
3. To address the health literacy and readability issue, OAC concepts are assigned consumer-friendly display (CFD) names. Terms are also assigned (consumer) familiarity scores. The validity of the CFD names and familiarity scores have been demonstrated in a few studies.
4. OAC contains a few hundred terms and concepts that are not present in UMLS and a few thousand different mappings between terms and concepts. (We are not providing an exact number here because both OAC and UMLS release new versions every few months.)
5. Future versions of OAC will contain semantic types and relations similar to, yet different from the UMLS.

The values in the OAC files are separated by tabs and are best viewed in Excel. Please note that in order to access these files, you will have to [sign the guest book](#). You can use the name and password you create when you sign the guestbook to access these files.

The Concepts Terms Flat File

The `concepts_terms_flat_file` contains terms and concepts and is similar to the `MRCONSO` file in UMLS. Each concept may have many terms that have mapped to it. Each of these terms is listed on a separate row, which means that there is more than one line associated with each concept. There is no particular order to the terms themselves, however. The first column is the concept ID (CUI) and the second column is the term. For each concept-term combination, there are flags indicating whether that term is the OAC preferred (or consumer-friendly display (CFD)) name, whether it is the UMLS preferred name, and whether it is disparaged. A disparaged term is one that is misspelled or has some other abnormality about it.

With each term is a column titled "status," which may contain the values "AMBIG" or "VAGUE". An ambiguous term is one that maps to more than one concept, and one or the other of these concepts is in mind. A vague term is one that maps to more than one concept, and it is not clear which one is in mind.

Each term also has a term score, which is a calculation of how easily understandable that term is. This number is derived from a regression model using the frequency of the term in several large text corpora. The cui score is a similar number which represents how understandable the concept is. It is derived from determining how closely related the concept is to known examples of easy and difficult concepts. The combo score attempts to combine the cui score and term score to arrive at another approximation of how understandable the cui/term is.

Another column is titled "reviewed". This simply refers to whether the mapping between the term and the concept has been manually reviewed. If it has not been reviewed, it may still be a valid mapping. If it has been reviewed, then it is definitely (in the opinion of the reviewers) a valid mapping.

Finally, a date column simply gives the date on which the flat file was generated.

The Ngrams Flat File

The ngrams flat file lists terms and phrases that have not mapped to the UMLS, but which, in the estimation of the reviewers, should map to medical concepts. The ngrams are not arranged in any particular order. For each ngram, a flag indicates whether it is meta, mod, disparaged, or misspelled. Every misspelled concept should be automatically disparaged. There is also a column which may contain a comment.

The Stop Concepts Flat File

The stop concepts flat file simply lists the CUIs and the names of concepts which we have judged to be excluded from the consumer health vocabulary. No term should map to any of these concepts.

The Incorrect Mappings Flat File

The incorrect mappings flat file lists combinations of CUIs and terms which are incorrect mappings. Of course many terms should not map to many concepts, but these are terms which actually have been mapped to these concepts under some system and which the reviewers have judged to be incorrect mappings. This list is not exhaustive. Many terms not listed next to a particular concept will also be incorrect mappings for that concept.