# 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 the
University of Utah, 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 has 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 mapping 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.

Description of Columns in Concept_Terms_Flatfile

<Column name>: <description>, <type>

CUI: UMLS Concept ID, String
Term: Term as found in text, String
CHV Preferred Name: Preferred name as defined in Consumer Health Vocabulary, String
UMLS Preferred Name: Preferred name for the CUI as defined by UMLS, String
Explanation: Explanation for the term (if available), String
UMLS preferred: A boolean variable (yes/no) describing whether UMLS prefers the term for the CUI, String
CHV preferred: A boolean variable (yes/no) describing whether CHV prefers the term for the CUI, String
Disparaged: A boolean variable (yes/no) indicating the term is misspelled or has some abnormality to it, String
Frequency Score: Estimate of the difficulty of the term based on its frequency in several large text corpora, Real
Context Score: Context based estimate of the difficulty of the term, Real
CUI Score: Estimate of the difficulty of the concept (CUI) derived from determining how closely related the concept is to known examples of easy and difficult concepts, Real
Combo Score: Combination of frequency, context and CUI scores (also uses whether or not the term is a top word), Real
Combo Score - No top words: A slight modification to Combo score that ignores top word criterion, Real
CHV String ID: Unique id for each entry in the CHV, String
CHV Concept ID: Unique id for each UMLS Concept represented in the CHV, String

NOTE: All score attributes have a range 0 to 1 (a higher score implies the term is easier). A value of -1 indicates the score could not be estimated.

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.