A Comparison of Meta-1 and HELP Terms: Implications for Clinical Data

Stanley M. Huff, M.D. and Homer R. Warner, M.D., Ph.D.
Department of Medical Informatics
LDS Hospital/University of Utah
Salt Lake City, UT 84143

Abstract

Terms from the HELP System's vocabulary were matched with Meta-1 terms on a word by word basis as well as on a phrase by phrase basis with the goal of exploring what steps might need to be taken if some future version (Meta-N) of the Metathesaurus were to be used to represent clinical data. Word by word matching revealed that 54% of HELP words were present in Meta-1, while 8% of HELP phrases had a corresponding phrase. The words that did not match in HELP were mostly adjectives and adverbs after taking into account misspellings and abbreviations. Phrase matches were low because of the inclusion of adjectives and adverbs in HELP clinical terms. If some future version of the Metathesaurus is to be used for representation of clinical data additional terms are needed as well as a grammar that permits construction of clinical phrases that include modifiers and time references.

Introduction

As described by Betsy L. Humphreys [1], "The Unified Medical Language System (UMLS) project is a major National Library of Medicine initiative designed to facilitate the retrieval and integration of information from many machine-readable sources, including descriptions of the biomedical literature, *clinical records* (italics ours), factual databanks, and medical knowledge bases." One major facet of the UMLS is the Metathesaurus, which is a set of canonical biomedical concepts (terms) with mappings to other biomedical vocabularies. While clinical records are clearly within the domain of the Metathesaurus, the priority in producing the first version of the Metathesaurus has focused on allowing indexing or referencing of clinical data rather than on actually representing clinical data. The NLM has been emphatic in stating that the Metathesaurus is not an attempt to impose a single medical vocabulary or a single standard record format on the biomedical community.

In spite of these facts, we have a special interest in the Metathesaurus as it pertains to representation of clinical data because of a long standing research interest in decision support systems for clinical users. As clinical researchers we would like to be able to exchange clinical knowledge bases, clinical research data, and clinical outcome data among institutions. One obstacle to the exchange of medical knowledge and medical data among institutions is the lack of a unified vocabulary or terminology as the basis for information exchange. While many vocabularies exist (CPT-4, SNOMED, ICD-9 CM, READ Codes, MeSH, ASTM Nomenclature [2,3,4,5,6,7]) none is entirely appropriate for the task because it is not comprehensive, it has no institutional support, it is not sufficiently detailed, it can not be updated in a timely fashion, it is unproven in the clinical environment, or it lacks the flexibility needed in clinical decision support processing. Within the HELP system [8] we have used our own hierarchical vocabulary (called PTXT, an acronym for Pointer to TeXT [9]) for over fifteen years and have developed a rich set of clinical terms within the system. But in the long term we see the need for a national or international vocabulary that has institutional support. While the first version of the Metathesaurus is not designed for representation of clinical data, we are hopeful that future versions might be.

Keeping the above facts in mind, and with the recent availability of the first version of the Metathesaurus (Meta-1), we have undertaken a comparison of PTXT and Meta-1 with the goal of exploring what steps might need to be taken if some future version (Meta-N) of the Metathesaurus were to be used to represent clinical data. We have compared the two vocabularies based on both a word by word content comparison and also on a multi-word or phrase based comparison.

Methods

Six relational files from Meta-1 were downloaded to a Charles River System computer which runs the UNOS operating system and supports the use of the UNIX word filter tools *awk* and *sed*. The specific files downloaded from Meta-1 and their total line counts are shown in table 1. The Meta-1 files were chosen to be the set of files with the richest

Table 1: Summary of files from Meta-1 used in the comparison.

File Name	Description	Size(lines)
m1.EC	Entry Combinations (MESH)	320
m1.LV	Lexical Variants	18527
m1.MC	Meta-1 Main Concepts	40009
m1.NT	Narrower Than (Candidate Synonyms)	2442
m1.RT	Related Terms	29098
m1.SY	Synonyms	21383
	total:	111779

word content but still related with a high degree of probability to a specific main concept. HELP system PTXT, on the other hand, is organized hierarchically into "data classes" which are roughly equivalent to clinical areas or departments. The data classes that were downloaded from PTXT and the total line counts are shown in table 2. The data classes chosen are those that account for more than 97% of the data stored on the HELP system, and all data classes that account for more than 1% of the clinical data. HELP System PTXT also contains SNOMED, ICD9-CM, and CPT-4 codes which are used for billing and medical records abstracting purposes. These codes were not included in the study since they are not used to store clinical data and the extent to which these vocabularies are included in Meta-1 has previously been described [10].

Table 2: Summary of data classes from HELP used in the comparison.

Description	Data Class	Size(lines)
Global Text Modifiers	0	556
Patient Demographic Data	1	177
EKG	3	2760
ICU Monitoring	4	407
ICU Monitoring	5	2317
Blood Gases	6	775
History and Physical Exam	7	7350
Pharmacy	8	7251
Clinical Chemistry	13	1372
Serology	14	921
Hematology	15	1749
Microbiology	16	2978
Toxicology	17	347
Radiology	20	6163
Nurse Care Plans	27	545
Nurse Charting	28	3764
Blood Bank	34	1496
Dietary, Food Services	35	2136
Respiratory Therapy	36	1786
Surgery Procedures	38	2558
Insurance Information	100	693
	total:	48101

After the files were downloaded they were processed as follows:

Common Processing:

1. Both PTXT and the Meta-1 files were converted to all uppercase characters using a "C" program written especially for this purpose.
2. The PTXT file was processed to remove leading and trailing blanks and to remove any redundant space characters. Abbreviations in the file were changed to a consistent format, i.e. "M.D." was changed to "MD", and numeric or special characters that were at the beginning of the text were removed. For example, typical lines in the PTXT input file look like this:

```
ROULEAUXFORMATIONOF
RBCS|P|0|127|13|0|0|236|0|122
ACANTHOCYTES
PRESENT|P|0|127|13|0|0|236|0|123
SCHISTOCYTES|P|0|127|13|0|0|236|0|124
TEARDROP CELLS|P|0|127|13|0|0|236|0|125
```

3. Each of the Meta-1 files was processed to remove numbers and special text from the first portion of the text field, and synonym markers were removed from the start of the text field and placed at the end of the line. The fields within a record were reordered so that the text came first, followed by a "U" to indicate that the text came from a UMLS source, followed by the sequence number of the item and then the file identifier. For example, the m1.SY file contents looked like this after initial processing:

```
TRICHLOROEPOXYPROPANE|U|39|SY||
CALCITRIOL|U|40|SY||
CALCITRIOL|U|41|SY||
CAFFEINE|U|42|SY||
```

Single Word Matching:

1. In preparation for comparison of the files on a word by word basis, the files described above were copied and then processed to break them into individual words. Each text phrase was broken at white space characters or punctuation and individual words were produced.
2. After breaking each phrase or term into individual words, the Meta-1 files were combined into one large file. All redundant words were discarded. The same process was repeated for the PTXT word file.
3. The sorted, unique words from PTXT and UMLS were then combined into a single file. Each term retained an identifying character so that its origin was clear. The combined file appeared as:

```
ABD|P
ABELSON|U
ABERRANCY|P
ABERRANT|P
ABERRANT|U
ABERRATIONS|U
ABETALIPOPROTEINEMIA|U
```

4. The combined file was then processed using a program that looked for nearest neighbor matches for each PTXT term in the file. Simple word stemming was also used so that "AMERICAN" was considered a match for "AMERICA" and "ACCIDENTS" was considered a match for "ACCIDENT", etc. In the above example, "ABERRANT|P" would match with "ABERRANT|U." Matched and unmatched words were counted and then saved for examination.
5. The accuracy of the matching algorithm was measured by S. Huff by manually reviewing 910 words which were not exact character by character matches.
6. The words which matched and did not match were reviewed and categorized as to part of speech and usage.

Whole Term or Phrase Matching:

1. The PTXT terms and Meta-1 terms were combined into a single large file. The resulting file appeared as shown below:

```
NAEGLERIA|U|30494|RT|
NAEGLERIA|P|16|1|65|2|1|5|0|0
NAEGLERIA FOWLERI|U|25225|SY|SNM/SY/E-4468|
NAEGLERIA GRUBERI|U|25225|SY|SNM/SY/E-4468|
NAESLLUNDII|P|16|1|61|3|9|1|3|0
NAF|P|16|13|1|16|0|59|232|0|1
NAF & CEPHA|P|16|13|1|16|0|59|232|0|3
NAF =|P|16|1|68|2|1|6|1|0|0|0
NAFCIL|U|25229|RT|MSH/RE/D009254
NAFCILLIN|U|34492|RT|
```

2. The combined file was processed using a nearest neighbor matching algorithm similar to the one described for the word by word match. In all cases a PTXT term is removed from further matching after it has been matched to a single Meta-1 term. In the above sample "NAEGLERIA|U" would be considered a match for "NAEGLERIA|P," while "NAESLLUNDII|P" would have no match. Where the context is known, abbreviations are considered a match to the full text of an item. Thus, "NAF|P" would match with "NAFCILLIN|U" because PTXT data class 16 type 131 terms are known to be antibiotics. "NAF & CEPHA|P" would not match in the example shown because it is a combined phrase and carries a different connotation than any of the single terms.
3. Additionally, the PTXT file was sent to Lexical Technology, Inc. to provide independent evaluation of the accuracy of the phrase matching.
4. The terms which matched and did not match were analyzed according to type of phrase and usage.

Results

The results of the Meta-1/PTXT comparison are summarized in table 3. The *Total Lines* column represents the total number of individual lines (terms or phrases) that the files contained before processing. The *Unique Lines* column represents the non-redundant lines remaining after processing, sorting, and redundant line removal. The *Total Words* column represents the number of non-unique word instances after lines (terms) were broken into individual words. The *Unique Words* column represents the number of unique words remaining after sorting and redundant word removal. *Word Matches* represents the number of PTXT unique words that had matches in the Meta-1 file. *Phrase Matches* is the number of unique PTXT phrases that had matches in the Meta-1 file. The *Word %Match* and *Phrase %Match* columns are simply the word match and phrase match counts converted to a percentage based on the total unique words and total unique lines in PTXT respectively. Note that the converse experiment (i.e. matching of Meta-1 phrases to PTXT) was not done. It should be clear that the numbers for the reverse match will not be identical to the PTXT matching statistics since in the reverse direction lexical variants may have non-unique mappings.

The accuracy of the matching algorithm for the word by word match was evaluated. Of the 15,838 words in PTXT 4,615 had exact character by character matches in Meta-1, while an additional 1,579 words matched based on lexical stemming. 2,801 words had less than 4 leading characters in common with any Meta-1 term and could not be matched by lexical stemming. The remaining 6,843 words could not be matched exactly by the stemming algorithm but had from 4 to 18 leading characters that matched a Meta-1 word. Of these words, 910 were manually reviewed by S. Huff to estimate the number of terms within this category that had matches in Meta-1. The results are shown in table 4 below. The number of true matches within this set of words may be different from the estimate made in table 4 by as much as 15%, which means that the total percent of matching words as shown in table 3 may be inaccurate by as much as 7% (in the range of 47% to 61%).

To evaluate the accuracy of the phrase matching algorithm the complete ASCII PTXT file was sent to Lexical Technology, Inc. Matching to Meta-1 phrases was then undertaken by them using the same lexical tools that are used to map Meta-1 terms to other vocabularies such as SNOMED, and ICD-9. The result showed 2,744 (8.2%) matching phrases using their programs as compared to 2,784 (8.4%) matching phrases obtained by our method [14].

Review of the phrases that matched showed that nearly all were nouns or noun phrases. Typically, matching phrases were names of procedures, names of compounds, names of diseases, or names of medical hardware. The non-matching words and phrases contained a high percentage of adjectives (high, low, increased, decreased, red, blood tinged, painful, intense, excruciating, left, right, inner, outer, upper, lower), adverbs (slowly, increasingly), and state of being verbs (is, was, have, will be). Other unmatched phrases included phrase like "of long duration," "with the presence of," "made worse by," "co-existing with," or had numeric quantifiers of date and time. The lack of non-nouns and non-noun phrases in Meta-1 is expected since Meta-1 was purposely limited to noun phrases for practical reasons on the first pass. More indepth study of the categories of word matches and misses and their quantitation is in progress.

Table 3: Summary of line counts and number of matches for Meta-1/PTXT comparison.

	Total Lines	Unique Lines	Total Words	Unique Words	Word Matches	Word %Match	Phrase Matches	Phrase %Match
Meta-1	111779	59458	223931	30630	------	--------	------	-------
PTXT	48101	33309	138418	15838	8581	54.2%	2784	8.4%

Table 4: Summary of manually reviewed words that could not be matched accurately by the matching algorithm.

Length of Match	Number of Terms	Number Reviewed	Number Matched	Percent Matched	Est. Matches In Category
4	1883	110	6	5	103
5	1486	110	18	16	243
6	907	110	27	25	223
7	1065	88	46	52	557
8	637	88	74	84	536
9	386	88	66	75	290
10	229	66	56	85	194
11-18	250	250	241	96	241
Total	6843	910	----	----	2387

From the 7,257 unmatched PTXT words, 368 were randomly selected and categorized as to the reason for not matching. The results are summarized in table 5 below. The "Odd" terms category included names of clinicians and other idiosyncrasies related to PTXT usage. Though the largest category of non-matching terms are simply not present in Meta-1, a significant number of errors are actually misspellings in PTXT or non-standard abbreviations.

Table 5. Distribution of errors in PTXT words that were not matched to any word in Meta-1.

Description	Count	%
Not Present	136	37
Lexical Variant	66	18
Abbreviation	66	18
Misspelled	47	13
Odd	40	11
Synonym	13	4
Total	368	100

Discussion

One somewhat surprising outcome is the small number of words that are actually used clinically within PTXT. Compared to the number of terms in a medical vocabulary like SNOMED, this small number of words is surprising. One explanation may be that the total number of medical terms is great when all of the eponyms and syndromes are included, but that the number of terms actually needed to communicate in the clinical setting is relatively small. If this is true, and the smaller set of terms can be identified, it would not be an insurmountable task to add these words into some future version of the Metathesaurus. If PTXT can be used as a pattern (assuming correction of the problems within PTXT of non-standard abbreviations and misspellings) roughly 3,000 additional words would need to be added to Meta-N to accommodate all of the clinical terms in PTXT.

However, consideration of the 8.4% phrase matching leads in a different direction. Clearly, Meta-1 would require additional terms (especially adjectives, adverbs) if it is to be used to represent day to day clinical data. Note that this is a very different statement than saying that Meta-1 is not useful for *referencing* clinical data. Though the study has not been done, it is the authors impression that the vast majority of HELP phrases could be indexed using Meta-1.

Simply adding words to Meta-1, however, does not mean that the clinical concepts used in the HELP system could be represented using the Meta-1 vocabulary. The words need to be combined in controlled ways or the meaning of an expression is lost. Many of the phrases in HELP are so specific ("Do you currently have a rash on your face?") that it would be hopeless to add terms without adding structure and philosophy. The addition of terms would need to be accompanied by a specific grammar to make it usable. This would be a major undertaking. Substantial institutional support would be required to create and maintain and update the vocabulary in a time frame acceptable to clinical users. This task is admittedly outside of the scope of the current UMLS project. Appropriately, the current focus is on creating a vocabulary for referencing biomedical literature and data rather than on representing clinical data. Whether a future project under the direction of the National Library of Medicine or perhaps in cooperation with some other governmental or private organization is feasible and desirable is a subject for general comment.

However, in our opinion, the benefits of a standard clinical vocabulary when combined with an appropriate data structure and grammar seem to be overwhelming. Such a vocabulary would permit the long anticipated electronic exchange of computable clinical data between medical institutions and government agencies, both for clinical research as well as outcome research studies. A unified vocabulary would also enable the practical exchange of executable clinical knowledge (like that available in HELP System frames) between institutions, which has been an unfulfilled promise for so many years. Research by our group and others [11, 12, 13] into frame structures and public domain clinical research databases as supported by the UMLS project is a step along the right path.

References

[1] Humphreys BL, Lindberg DAB. Building the Unified Medical Language System.
In: Kingsland LC, ed, Proceedings of 13th annual symposium on computer applications in medical care. New York: Institute of Electrical and Electronics Engineers, 1989; 475-80.

[2] Clauser SB, Fanta CM, Finkel AJ. ed. CPT-1984: Physicians Current Procedural Terminology. American Medical Assn. Chicago, 1984.

[3] Cote RA, ed. Systematized Nomenclature of Medicine. College of American Pathologists, Skokie, IL.

[4] United States National Center for Health Statistics: International Classification of Diseases, 9th Revision, Clinical Modifications. Ann Arbor, MI.

[5] Read JD, Benson TJR. Comprehensive Coding. British Journal of Healthcare Computing, Vol 3, No 2, 22-25, may 1986.

[6] National Library of Medicine, Medical Subject Headings Section, Medical Subject Headings, Annotated Alphabetical List, 1989, Bethesda, MD.

[7] Gabrieli ER. Nosologic standards for construction of a new biomedical nomenclature. Draft document ASTM, Philadelphia, PA, 1987.

[8] Pryor TA, Gardner RM, Clayton PD, Warner HR. The HELP system. J Med Systems 1983; 7:87-102.

[9] Warner HR. Computer assisted medical decision-making. Academic Press Inc., New York, New York, 1979.

[10] Tuttle M, Sheretz D, Erlbaum M, Olson N, Nelson S. Implementing Meta-1: The First Version of the UMLS Metathesaurus. In: Kingsland LC, ed, Proceedings of 13th annual symposium on computer applications in medical care. New York: Institute of Electrical and Electronics Engineers, 1989; 483-87.

[11] Miller RA, Masarie FE, Claudon CH, Giuse NB, Warner HR, Bouhaddou O. Mapping of Medical Knowledge Representations: INTERNIST-I, HELP, and MESH. Final Task Report, Task 4, Unified Medical Language System, Contract No. N01-LM-6-3522, 1987 October 9.

[12] Masarie FE, Cimino JJ, Giuse NB, Miller RA. Mapping Between Controlled Vocabularies: QMR and DXplain, Report of Results of Task 5, Subtask 6, N01-LM-6-3522, 1988 April 14.

[13] Cimino JJ, Barnett GO. Automated translation between medical terminologies using semantic definitions. In: 1989 AAMSI Proceedings.

[14] Tuttle M. Personal Communication, May 1990. Lexical Technology, Inc.