The UMLS-CORE Project – A Study of the Problem List Vocabularies Used in Large Health Care Institutions

Kin Wah Fung, MD MS MA, Suresh Srinivasan, MS

National Library of Medicine, Bethesda, MD, USA

Correspondence and reprints:

Kin Wah Fung
Building 38A, Rm9S914, MSC-3826
National Library of Medicine
8600 Rockville Pike
Bethesda, MD 20894

Telephone: 301 – 435 3151
Fax: 301 – 496 0663
Email: kwfung@nlm.nih.gov

Abstract

Objective: To study existing problem list vocabularies, and to identify a CORE (Clinical Observations Recording and Encoding) Subset of UMLS (Unified Medical Language System) concepts that occur frequently in problem list data.

Methods: Problem list terms and their usage frequencies were collected from large health care institutions. The pattern of usage of the terms was analyzed. The local terms were mapped to the UMLS. Based on the mapped UMLS concepts, the degree of overlap between the problem list vocabularies was analyzed.

Results: Six institutions submitted 76,237 terms and their usage frequencies in 14 million patients. The distribution of usage was highly skewed. 10% of unique terms already covered over 85% of usage. The most frequently used 14,395 terms, covering 95% of usage in each institution, were exhaustively mapped to the UMLS. 13,261 terms (92%) were successfully mapped to 6,776 UMLS concepts. This subset of UMLS concepts constituted the UMLS-CORE Subset. Less frequently used terms were generally less mappable to the UMLS. The average pairwise overlap of the problem list vocabularies was only 21% (median 19%). Concepts that were shared among institutions were used eight times more frequently than concepts that were unique to one institution.

Conclusion: Most of the frequently used problem list terms could be found in standard terminologies. The overlap between existing problem list vocabularies was low. The UMLS-CORE Subset identified in this study can be used as a starter set to create problem list vocabularies. This will save developmental effort, reduce variability of problem list vocabularies and enhance interoperability of problem list data.

Keywords: Unified Medical Language System, problem-oriented medical record, problem list, Electronic Health Record, SNOMED Clinical Terms, ICD-9-CM, controlled medical terminology, medical vocabulary

Introduction

The problem-oriented approach of organizing information in a medical record was first advocated by Weed almost 40 years ago.[1] Central to this approach is a problem list which is "a complete list of all the patient's problems, including both clearly established diagnoses and all other unexplained findings that are not yet clear manifestations of a specific diagnosis, such as abnormal physical findings or symptoms". According to Weed, this list should also cover "psychiatric, social and demographic problems". Though the adoption of Weed's problem-oriented approach to the whole medical record has been limited, the use of problem lists is widespread in both paper and computer based medical records. Furthermore, many sanctioning bodies and medical information standards organizations consider the problem list as an important element of the Electronic Health Record (EHR), including the Institute of Medicine, Joint Commission, American Society for Testing and Materials and Health Level Seven. [2-6] An encoded problem list is also one of the requirements in the Interim Final Rule for the meaningful use of EHR published by the Department of Health and Human Services. [7] In a recent national survey on the use of the EHR in U.S. hospitals, an expert panel considers the problem list an essential component of both a basic and comprehensive EHR. [8]

Problem lists have value beyond clinical documentation. Other common uses include the generation of billing codes and clinical decision support. To drive many of these functions, an encoded problem list (as opposed to data entered as free-text) is often required. This probably explains why the problem list is often the first (if not the only) location within an EHR with encoded clinical statements. This paper explores the use of controlled vocabularies in electronic problem lists and their associated problems.

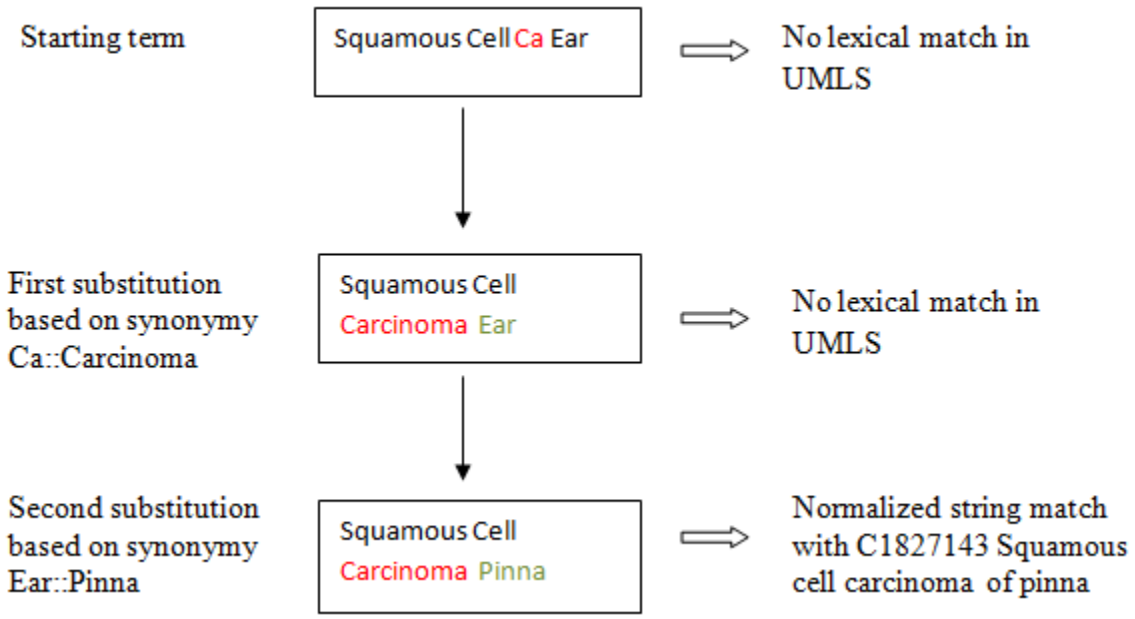
In an ideal world, everybody should use a single, standardized problem list vocabulary. In reality, most institutions use their own local vocabularies. In the U.S., even though SNOMED CT is the designated vocabulary for problem lists by the Consolidated Health Informatics Initiative, [9] the adoption of SNOMED CT has not been widespread. [10] We started the UMLS-CORE Project in 2007 to study problem list vocabularies. As the flagship terminology product of the U. S. National Library of Medicine (NLM), the Unified Medical Language System (UMLS) is a valuable resource for terminology research. [11] CORE stands for **C**linical **O**bservations **R**ecording and **E**ncoding, a mnemonic referring to the capture and codification of clinical information in the summary segments of the EHR such as the problem list, discharge diagnosis and reason for encounter. The UMLS-CORE Project has two goals:

1. To study and characterize the problem list vocabularies of large health care institutions in terms of their size, pattern of usage, mappability to standard terminologies and extent of overlap.
2. To identify a subset of UMLS concepts that occur with high frequency in problem lists to facilitate the standardization of problem list vocabularies.

Methods

We asked large scale health care institutions to submit their problem list terms together with the actual frequency of usage in their clinical database. We also requested any mapping from the local problem list terms to standard terminologies (e.g. ICD-9-CM, SNOMED CT) if available. We mapped the local terms to the UMLS using the 2008AA release. We only used exact maps that captured the full meaning of the local terms (i.e. no inexact or broader/narrower maps). UMLS mapping was done in three sequential steps. The first step was lexical matching. We first

looked for exact (case-insensitive) and normalized matches using all English terms in the UMLS. Normalization was done using the 'norm' function of the UMLS lexical tools, in order to abstract away from differences due to word inflection, case and word order. [12] For example, 'Perforating duodenal ulcers' and 'Duodenal ulcer, perforated' both normalize to the same string 'duodenal perforate ulcer'. We further enhanced lexical matching by synonymous word/phrase substitution. [13] Local terms that did not have exact or normalized matches were parsed for words or phrases that occurred in our internal word/phrase synonymy table (used in UMLS quality assurance). If found, we substituted the word/phrase with the alternative word/phrase and repeated the matching. We allowed a maximum of two word/phrase substitutions to avoid unintentional meaning drift. Here is an example of lexical matching with synonymous word substitution:



Terms not mapped by lexical matching were passed through the second step, which made use of local maps to standard terminologies (either ICD-9-CM or SNOMED CT) if they were available.

Not all local maps were exact maps. Only those maps that were explicitly labeled as exact maps by the institutions were used to map automatically (i.e. without manual review) to the UMLS. For example, the local term 'Adrenal insufficiency' was mapped to the SNOMED CT concept '111563005 Adrenal hypofunction' by one institution and it was labeled as an exact map. This map was used to map the local term to the UMLS concept C0001623 containing that SNOMED CT concept. Local maps that were not labeled explicitly as exact maps were manually reviewed and only those that were considered exact maps were used.

The final step was manual mapping. All terms that remained unmapped after the first two steps were manually mapped to the UMLS, using the RRF browser included in the UMLS as the searching tool. [14] All terms that were ultimately unmapped were analyzed for reasons of failure.

All subsequent analysis was based on those terms that could be mapped to the UMLS. We used the UMLS concept (identified by its Concept Unique Identifier, or CUI) as the proxy for the local term. We calculated the pairwise overlap between institutions as follows:

$$\text{Percent of overlap of A and B} = \frac{\text{\# of CUIs common to A and B}}{\text{Total \# of Unique CUIs in A and B}} \times 100\%$$

We also analyzed the relationship between the level of sharing of concepts and their usage.

Results

A. Characteristics of the institutions and datasets

We obtained datasets from the following six health care institutions: Kaiser Permanente (KP), Mayo Clinic (MA), Intermountain Healthcare (IH), Regenstrief Institute (RI), University of Nebraska Medical Center (NU) and Hong Kong Hospital Authority (HA). HA was the only institution outside the U.S. All the U.S. datasets covered both ambulatory care and hospital patients. The HA dataset represented the discharge diagnoses of hospital inpatients. The characteristics of these institutions and their datasets are summarized in Table 1.

The institutions were of relatively large scale and provided service in all major medical specialties including Internal Medicine, General Surgery, Pediatrics, Obstetrics & Gynecology, Psychiatry and Orthopedics. The problem lists in these institutions were generated and maintained by physicians as part of the care process, and the information was also available to other caregivers. All institutions were using or planning to use the encoded problem list data for purposes beyond clinical documentation. These uses included: clinical decision support (e.g. pharmacy alerts, best practice alerts, suggestion of order sets), generation of administrative codes (e.g. ICD-9-CM codes and DRG for billing, ICD-10 codes for public health reporting), clinical research (e.g. identification of study subjects, enrollment to research protocols) and compilation of management statistics.

The six datasets together covered 14 million patients. For HA and MA, the datasets represented all patients encountered by the system in three years. The RI data were collected over one year. For IH, KP and NU, the datasets included all patients currently registered in their systems. Most of the usage data were patient-based, meaning that a given problem in one patient would be counted only once, though it could be documented multiple times in different encounters. Only

the MA usage data were encounter-based. The average number of problems per patient varied from three to seven. The size of the problem list vocabularies varied considerably across institutions, ranging from just over 3,000 to almost 27,000 unique terms.

B. Usage pattern of terms

For each institution, we looked at the statistical distribution of usage. Figure 1 shows the percentage of unique terms (vertical axis) required in each institution to cover a certain percentage of total usage (horizontal axis). We used the percentage (instead of the absolute number) of unique terms to facilitate cross-institution comparison, because there was significant variation in the size of the problem list vocabularies. 10% of unique terms already covered more than 85% of usage in all institutions except RI. To cover 95% of usage, the average percentage of terms required was 21%. The same skewed distribution, i.e. a small proportion of highly-used terms and a long tail of rarely used one, was found in all institutions

C. Mapping to the UMLS

We performed algorithmic lexical mapping for all 76,237 terms in the six datasets. For pragmatic reasons, we only did full mapping (including manual mapping) for the most frequently used 14,395 terms that covered 95% of usage in each institution. The following analysis and percentages were based on this subset of terms.

1) Results of UMLS mapping

Lexical mapping yielded maps for 10,812 terms (75%). Among them, exact string match (case-insensitive) found 8,102 terms (56%), normalized string match found 2,035 terms (14%) and synonymous word/phrase substitution found an additional 675 terms (5%). The next mapping

step made use of local maps when available. Some mapping to ICD-9-CM was available for HA, MA and RI, while mapping to SNOMED CT was available for IH, KP and NU. Only maps from HA, IH and KP were explicitly labeled for the degree of exactness. Altogether, these local maps yielded UMLS maps for 1,007 terms (7%). We manually reviewed the remaining 2,576 terms and mapped 1,442 (10%) terms, leaving 1,134 terms (8%) ultimately unmapped. The mapping results are summarized in Table 2. Altogether, 13,261 local terms were mapped to 6,776 UMLS concepts, constituting what we called the UMLS-CORE Subset. This Subset represents the UMLS concepts most commonly used in problem lists.

2) Reasons for unmappability

We assigned each unmapped term to one of eight categories which were derived empirically (Table 3). The commonest category (53%) was terms that included a high level of detail. One example was 'Benign prostatic hyperplasia with age related prostate cancer risk and obstruction'. While 'Benign prostatic hyperplasia' and 'Benign prostatic hyperplasia with obstruction' both existed in the UMLS, the further qualification with cancer risk was not found. Many of these terms could be considered subtypes of existing terms with additional specificity. On the other hand, 11% of the unmapped terms were very general e.g. 'Abnormal blood finding'. 7% of terms conveyed administrative rather than clinical information e.g. 'Other Mr # exists', presumably to alert the healthcare provider of another patient with the same name. Another 7% of terms contained laterality information, which was not usually captured in standard terminologies. Again, like the highly specific terms, these terms could be considered subtypes of existing terms. Terms conveying negative findings and composite concepts each made up 3% of unmapped terms. A small number of terms (2%) were ambiguous and could not be mapped without clarifying their meaning. For example, red conjunctiva can refer to either acute conjunctivitis or

conjunctival hemorrhage which are distinct clinical entities. Finally, there were 13% of miscellaneous terms not easily classifiable to the other categories. For example, while 'Sinusitis', 'Acute sinusitis' and 'Chronic sinusitis' existed in the UMLS, 'Subacute sinusitis' was not found.

3) Relationship of usage to mapping

To study the relationship between frequency of usage and mappability to the UMLS, we arranged all 14,395 terms in descending order of usage. For each quantile, we calculated the percent of unmapped terms. In figure 2, the horizontal axis shows the percentiles (e.g. 10% means the most frequently used 10% of terms) and the vertical axis shows the percent of unmapped terms among the terms within that percentile. More frequently used terms were generally more mappable. As we included more of the less frequently used terms, the percentage of unmapped terms increased. This finding is understandable because frequently used terms are more likely to have made their way into standard terminologies and thus the UMLS.

4) Overlap between institutions

We calculated pairwise overlap between problem list vocabularies based on the terms that could be mapped to the UMLS, using the CUI as the basis for comparison. The pairwise overlap showed considerable variability. (Table 4) The lowest pairwise overlap was between HA and RI (11%) and the highest between KP and NU (29%). HA had the lowest mean pairwise overlap (15%) with all other institutions, while NU had the highest (24%). The overall mean pairwise overlap for all six institutions was 21% (median 19%).

Of the 6,776 concepts (CUIs), 4,201 concepts (62%) were unique to only one institution while 2,575 concepts (38%) were shared. When concept sharing was correlated with usage data, the unique concepts were used much less frequently than the shared ones. Even though unique

concepts made up 62% of total concepts, they accounted for only 15% of usage (the mean usage over all six institutions). On the other hand, the 38% of shared concepts accounted for 77% of usage. Put in another way, each shared concept was used eight times more frequently on average than a unique concept.

Discussion

Despite the efforts to standardize medical terminologies, problem list vocabularies are still very much products of independent creation and evolution. In the beginning, institutions created their own vocabularies, which could be derived from some pre-existing term lists or mined from clinical data. [15-17] The subsequent changes of these vocabularies are driven by a multitude of factors related to the various uses of the problem list data. The primary use of the problem list is clinical documentation. The problem list is "a table of content and an index" which provides a convenient summary of the patient's problems and significant co-morbidities. This information helps to facilitate the continuity of care, formulation of plan of treatment or further investigations and management of risk factors. To facilitate clinical documentation, the terms in the vocabulary need to resemble closely clinical parlance so that users can find their terms easily. Most institutions allow users to request terms that they cannot find. These requests are usually vetted through an editorial process and added to the vocabulary if necessary. [18] Factors like patient mix, level of care (primary or specialty care), care setting (e.g. inpatient vs. ambulatory), specialty distribution, user preferences, term request process and editorial policy all affect the scope and granularity of the problem list vocabulary.

Problem list data are often used to drive functions other than clinical documentation e.g. generation of billing codes, supporting clinical research and quality assurance. The type and extent of these additional uses vary significantly across institutions, [19] and they pose their own requirements on the problem list vocabulary. It is not uncommon that different requirements are in conflict (e.g. clinical documentation vs. billing requirements) and the result is often a compromise between the usability of the problem list vocabulary, primarily for the purpose of clinical documentation, and the usefulness of the data captured for other purposes.

This 'letting a thousand flowers bloom' scenario has two problems. Firstly, every institution that creates its own problem list vocabulary is duplicating work that has been done by others. Given that there may be special needs unique to a particular institution, many uses and requirements of the problem list vocabulary are common and a lot of the duplicated efforts could be avoided. Moreover, from the data interoperability perspective, the lack of a common set of problem list terms is an impediment to data sharing.

Our study shows that the average pairwise overlap of existing problem list vocabularies is a meager 21%, hardly encouraging for the sharing of problem list data. One good news is that actual usage is concentrated heavily on relatively few terms, which makes the problem more tractable because we only have to standardize a relatively small proportion of terms to reap large benefits in data interoperability. Another encouraging finding is that terms that are shared are the ones that are heavily used. This reduces the effort and increases the yield of standardization.

We think the UMLS-CORE Subset identified in this study can help to alleviate the above problems. We recommend the use of the CORE Subset as a 'starter set' for problem list vocabularies. We believe that the datasets from which this subset is derived are large and broad enough to cover most data entry requirements in healthcare institutions providing service in all major medical specialties. This will save considerable time and effort compared to starting from scratch. The CORE Subset can enhance data interoperability in two ways. Firstly, by pre-selecting a set of terms, one can avoid the arbitrary (unintentional) variations in the creation of local vocabularies that are not based on clinical necessity or importance. For example, one could choose either of the two terms 'Infectious colitis, enteritis, and gastroenteritis' or 'Infectious gastroenteritis' (both are actual terms from standard clinical terminologies) to describe a patient's illness. Semantically, the two terms can be considered different as the former explicitly states 'colitis' while the latter does not. However, in most clinical situations, the diagnosis of infectious gastroenteritis is made without the need for radiological or endoscopic studies. Whether the colon is involved is usually neither known nor clinically important, because it does not affect the patient's management or prognosis. Limiting the choice to either one term (but not both) will reduce unnecessary variation in problem list data. A further way in which the CORE Subset can enhance data interoperability is that if existing problem list terms can be mapped to concepts in the CORE Subset, they can become the lingua franca for data exchange.

Among the most frequently used 14,395 terms, 92% could be found in standard terminologies in the UMLS. It is encouraging to see that existing terminologies already cover the majority of the frequently used terms in problem lists. There was a previous study, also by NLM, called the Large Scale Vocabulary Test (LSVT) Study, which evaluated the extent to which existing

medical terminologies covered the terms needed for health information systems. [20] It is interesting to compare the results of the LSVT with the present study. The LSVT found that only 64% of terms used in problem lists were covered by the medical terminologies in the UMLS. The lower percentage in LSVT could be explained by two reasons. Firstly, the LSVT mapped all submitted terms while the present study only focused on the frequently used ones. As we have shown, more frequently used terms were more likely to be found in the UMLS, and therefore a higher percentage of mappable terms were found in the present study. Secondly, the UMLS used in LSVT was considerably smaller, containing about 250,000 concepts and 600,000 terms compared to 1.5 million concepts and 6.4 million terms in the 2008AA version used in this study. This has no doubt contributed to the higher coverage in the present study.

One important finding which is common in both LSVT and the present study is that a significant proportion of terms that did not exist in standard terminologies could be derived from existing terms by the addition of modifiers. In LSVT, two-thirds of the terms that did not have exact matches were narrower in meaning than an existing concept in the controlled vocabularies, and most of these terms could be represented with the addition of modifiers to the broader concept. In the present study, more than half of the terms that were not found in the UMLS were highly specific concepts or contained laterality information (Table 3). Many of these terms could be represented by adding modifiers to existing concepts. The implication is that post-coordination (i.e. the creation of new meaning by combining existing concepts to modify their meaning) will be a good way to 'fill in the gaps', i.e. to provide the missing concepts requested by users. The advantage of post-coordination (compared to just adding completely new terms de novo) is twofold. Firstly, if everybody follows the same rules to combine concepts to generate new

meanings, at least theoretically one can determine computationally whether any two new concepts are equivalent even if they are created independently. Secondly, the new concepts will maintain their links to existing concepts. For example, if the new concept 'Left kidney stone' is created by adding the qualifier 'Left' to the existing concept 'Kidney stone' in the CORE Subset, the system will be able to recognize that the new concept is a subtype of kidney stone, and meaningful aggregation of existing and new concepts can still occur.

Conclusion

The problem list has been widely embraced as an efficient way to organize clinical information in EHRs. Problem list information that is encoded is required to invoke many of the intelligent functions of an EHR. The lack of a publicly available set of problem list terms results in duplication of effort in vocabulary development and impaired data interoperability. A high proportion of commonly used problem list terms are already found in standard terminologies. There is only modest overlap between existing problem list vocabularies from large health care institutions. The terms that are shared are used more heavily than terms that are not shared. A CORE Subset of UMLS concepts which covers the most frequently used problem list terms is identified. The use of the CORE Subset will save effort in vocabulary development, reduce variability and facilitate sharing of problem list data.

Acknowledgements

The authors would like to thank Christopher Chute, Robert Dolin, Stanley Huff, Vicky Fung, James Campbell, Naveen Maram, and Jeff Warvel for providing their institution's datasets for this study. Thanks also to John Kilbourne for the help in reviewing the UMLS mappings. Special thanks to Clement McDonald for providing insightful comments on the manuscript.

References:

1. Weed LL. The problem oriented record as a basic tool in medical education, patient care and clinical research. Ann Clin Res 1971;3:131-4.
2. Dick RS, Steen EB, Detmer DE. The computer-based patient record: an essential technology for health care, revised ed. National Academy Press, Washington, DC, 1997.
3. Committee on Data Standards for Patient Safety, Board on Health Care Services, Institute of Medicine. Key Capabilities of an Electronic Health Record System: Letter Report, 2003.
4. Hospital accreditation standards: accreditation policies, standards, elements of performance (HAS): Joint Commission on the Accreditation of Healthcare Organizations, 2009.
5. ASTM Standard E2369 - 05e1 "Standard Specification for Continuity of Care Record (CCR)" ASTM International, West Conshohocken, PA, DOI: 10.1520/E2369-05E01: ASTM, 2005.
6. Dickinson G, Fischetti L, Heard S, (Eds). HL7 EHR System Functional Model - Draft Standard for Trial Use: HL7, July 2004.
7. Department of Health and Human Services. Health Information Technology: Initial Set of Standards, Implementation Specifications, and Certification Criteria for Electronic Health Record Technology; Interim Final Rule, 2010.
8. Jha AK, DesRoches CM, Campbell EG, Donelan K, Rao SR, Ferris TG, Shields A, Rosenbaum S, Blumenthal D. Use of electronic health records in U.S. hospitals. N Engl J Med 2009;360:1628-38.
9. Department of Health and Human Services. Consolidated Health Informatics (CHI) Initiative; Health Care and Vocabulary Standards for Use in Federal Health Information Technology Systems: Federal Register, Dec 2005.
10. Giannangelo K, Fenton SH. SNOMED CT survey: an assessment of implementation in EMR/EHR applications. Perspect Health Inf Manag 2008;5:7.
11. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med 1993;32:281-91.
12. McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. Proc Annu Symp Comput Appl Med Care 1994:235-9.
13. Hole WT, Srinivasan S. Discovering missed synonymy in a large concept-oriented Metathesaurus. Proc AMIA Symp 2000:354-8.
14. UMLS RRF Browser http://www.nlm.nih.gov/research/umls/implementation_resources/metamorphosys/RRF_Browser.html.
15. Chute CG, Elkin PL, Fenton SH, Atkin GE. A clinical terminology in the post modern era: pragmatic problem list development. Proc AMIA Symp 1998:795-9.
16. Elkin PL, Mohr DN, Tuttle MS, Cole WG, Atkin GE, Keck K, Fisk TB, Kaihoi BH, Lee KE, Higgins MC, Suermondt HJ, Olson N, Claus PL, Carpenter PC, Chute CG. Standardized problem list generation, utilizing the Mayo canonical vocabulary embedded within the Unified Medical Language System. Proc AMIA Annu Fall Symp 1997:500-4.
17. Brown SH, Miller RA, Camp HN, Guise DA, Walker HK. Empirical derivation of an electronic clinically useful problem statement system. Ann Intern Med 1999;131:117-26.

18. Warren JJ, Collins J, Sorrentino C, Campbell JR. Just-in-time coding of the problem list in a clinical environment. Proc AMIA Symp 1998:280-4.
19. Wang SJ, Bates DW, Chueh HC, Karson AS, Maviglia SM, Greim JA, Frost JP, Kuperman GJ. Automated coded ambulatory problem lists: evaluation of a vocabulary and a data entry tool. Int J Med Inform 2003;72:17-28.
20. Humphreys BL, McCray AT, Cheh ML. Evaluating the coverage of controlled health data terminologies: report on the results of the NLM/AHCPR large scale vocabulary test. J Am Med Inform Assoc 1997;4:484-500.

	HA	IH	KP	MA	NU	RI
Type of service	inpatient	mixed	mixed	mixed	mixed	mixed
Inpatient percentage	100%	10%	No data	20%	15%	50%
Patient count (million)	1.3	0.36	10	1.5	0.5	0.16
Period of data retrieval	3 years	all current patients	all current patients	3 years	all current patients	1 year
Nature of problem data	patient-based	patient-based	patient-based	encounter-based	patient-based	patient-based
Total problem count (million)	4.1	1.1	52	10	2.7	0.66
Average problem per patient	3.1	3.0	5.2	6.8	5.3	4.2
Total unique terms	12,449	5,685	26,890	14,921	13,126	3,166
Unique terms covering 95% of total usage	2,635	1,077	2,961	3,610	3,320	792

Table 1. Characteristics of the institutions and their datasets (HA - Hong Kong Hospital Authority, IH - Intermountain Healthcare, KP - Kaiser Permanente, MA - Mayo Clinic, NU - University of Nebraska Medical Center, RI - Regenstrief Institute)

	Number of terms	**Percent**
Lexical matching	10,812	75%
Mapping based on local maps provided by sources	1,007	7%
Manual mapping	1,442	10%
Terms did not exist in UMLS	1,134	8%
Total	14,395	100%

Table 2. Mapping to the UMLS

Category	**%**	**Example**
Highly specific	53%	Benign prostatic hyperplasia with age related prostate cancer risk and obstruction
Very general	11%	Abnormal blood finding
Administrative	7%	Other Mr # exists
Laterality	7%	Renal stone, right
Negative finding	3%	No urethral stricture
Composite concept	3%	Diarrhea with dehydration
Meaning unclear	2%	Conjunctiva red
Miscellaneous	13%	Subacute sinusitis

Table 3. Categorization of terms that could not be mapped to the UMLS

Institution	Pairwise overlap with						
	HA	IH	KP	MA	NU	RI	Mean overlap
HA		13%	17%	17%	18%	11%	15%
IH	13%		25%	19%	25%	27%	22%
KP	17%	25%		29%	29%	17%	23%
MA	17%	19%	29%		31%	14%	22%
NU	18%	25%	29%	31%		19%	24%
RI	11%	27%	17%	14%	19%		18%

Table 4. Pairwise overlap between the problem list vocabularies

CUI appearing in	Number of CUIs	Percent of total CUI	Corresponding Mean Usage
1 dataset	4,201	62%	15%
2 datasets	1,130	17%	9%
3 datasets	607	9%	9%
4 datasets	391	6%	11%
5 datasets	282	4%	17%
6 datasets	165	2%	30%
Total	6,776	100%	92%

Table 5. Distribution of CUIs among datasets and the corresponding usage coverage

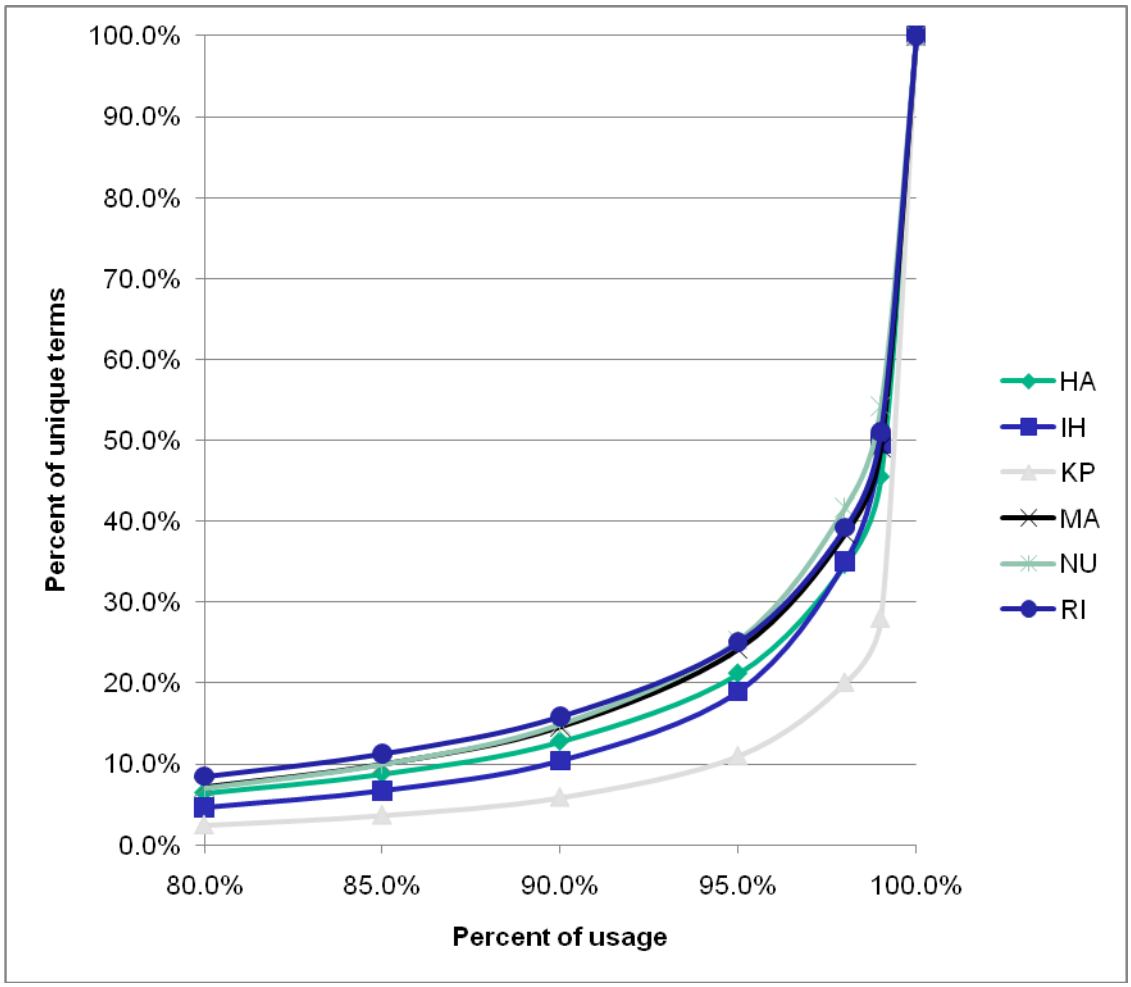


Figure 1. Usage pattern of the problem list terms (HA - Hong Kong Hospital Authority, IH - Intermountain Healthcare, KP - Kaiser Permanente, MA - Mayo Clinic, NU - University of Nebraska Medical Center, RI - Regenstrief Institute)

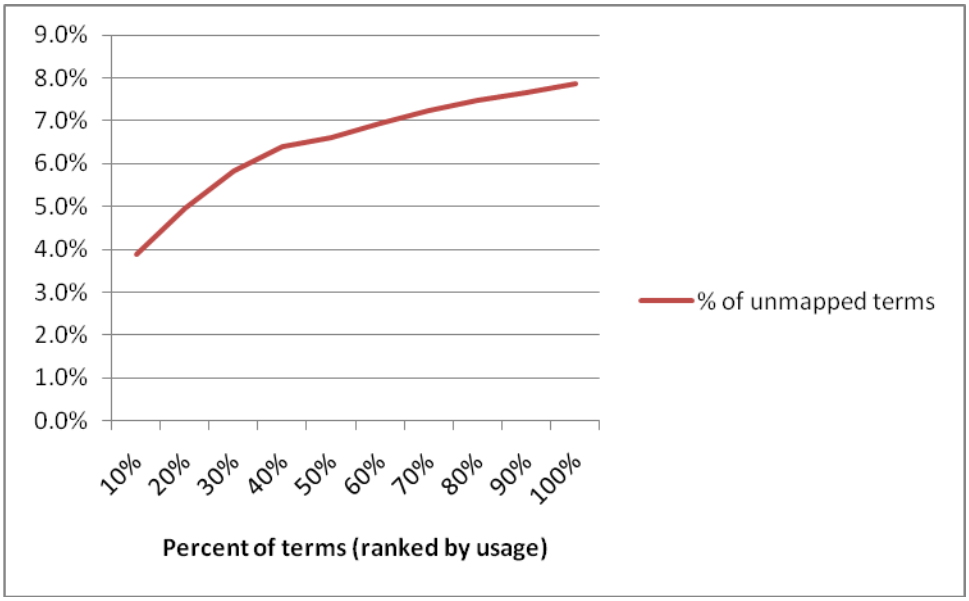


Figure 2. Relationship of term usage to mappability to the UMLS