Do Existing Controlled Vocabularies Contain Terminology Needed for Patient Records?

by Betsy L. Humphreys, MLS

A 1996-97 study conducted by the National Library of Medicine and the Agency for Health Care Policy and Research aimed to determine how well existing vocabularies cover the concepts and terms needed in health information systems. The author describes the hypothesis underlying the study, how the process worked, its findings, and implications for future vocabulary development.

From August 1996 to January 1997, 63 health professionals and researchers in the US and Canada submitted more than 41,000 terms to test the combined coverage of a group of existing controlled health terminologies. Sponsored by the National Library of Medicine (NLM) and the Agency for Health Care Policy and Research (AHCPR), this Large-Scale Vocabulary Test (LSVT) was designed to determine how well existing terminologies, taken as a set, cover the concepts and terms needed in health information systems, particularly in clinical data systems. The goal was to obtain data that would be useful in the development of federal policy on health data standardization and also help NLM to set priorities for expansion of the Unified Medical Language System[1] (UMLS) Metathesaurus.[2]

Planning for the LSVT predated the passage of the Health Insurance Portability and Accountability Act of 1996 (HIPAA), but the results are arriving at an opportune time to assist the Department of Health and Human Services (HHS) and the National Committee on Vital and Health Statistics (NCVHS) in carrying out the law's administrative simplification provisions. These provisions require that HHS, in consultation with the NCVHS and private-sector standards groups, establish standards for administrative health transactions, including the code sets to be used in such transactions. More germane to the results of the LSVT is the additional requirement that HHS and NCVHS submit recommendations to Congress in 2000 about any actions that should be taken to promote the development of full electronic medical records.

A detailed description of the LSVT methodology and the results of the initial analyses performed on the data appear in the November/December 1997 issue of the Journal of the American Medical Informatics Association.[3] The actual test data are available to UMLS users for further analysis and research.[2] What follows is an overview of the hypotheses and unique features of the LSVT, a summary of its methods and key results, and a brief discussion of potential implications for health data standardization and the development of computer-based patient record systems.

Hypotheses and Unique Features

The LSVT was designed to test two hypotheses:

1. Taken as a set, existing controlled health vocabularies contain most of the concepts and terms needed for health information systems, including detailed clinical information systems
2. The Internet and the UMLS resources can support a national experiment to assess the combined coverage of existing health vocabularies

The test differed from previous studies of health vocabularies[4,5] in several ways. As indicated previously, it assessed the aggregate coverage of many existing health vocabularies. Most previous studies have focused on comparing the coverage of individual controlled vocabularies or classifications. The LSVT involved widely dispersed participants, each using the same World Wide Web application. The ability to submit terms from any location with a suitable Internet connection led to a broad and diverse group of participants. The set of terms searched in the test was defined by the participants' interests, since the terms used were needed or desired for real tasks at the participants' sites. In other vocabulary tests, the terms searched were

extracted solely from samples of clinical text in patient charts. The LSVT allowed participants to search all the existing controlled vocabularies through a single search mechanism, which applied the same robust search capabilities and algorithms to each vocabulary. This avoided the potentially confounding effect of different search interfaces, which had been raised as a criticism of earlier studies.

Vocabulary Test Set

The LSVT test set included some 30 vocabularies and classifications present in the 1996 edition of the UMLS Metathesaurus, plus others subsequently added to the 1997 and 1998 editions of the Metathesaurus. There were large systems with broad subject coverage, (e.g., SNOMED International, the Read Codes, ICD-9-CM, the Medical Subject Headings [MeSH]), and more focused vocabularies and classifications such as Nursing Interventions Classification, Home Health Care Classification, Universal Medical Device Nomenclature System, and Logical Observations Identifiers Names and Codes (LOINC). Vocabularies developed for use in adverse drug event reporting (e.g., COSTART), in ambulatory record systems, (e.g., Computer-Stored Ambulatory Record System [COSTAR]), and in expert or other knowledge-based systems (e.g., DXplain, AI/RHEUM, PDQ/CancerNet) were also represented.

Methods

NLM and AHCPR encouraged anyone with a good Internet connection and a real task that required controlled health vocabulary to submit terms to the test.[6] Use of the Internet, the UMLS Knowledge Source Server, and a special World Wide Web application allowed the collection of comparable data from many different locations and environments.[7] Participants were asked to categorize the groups of terms they submitted by such characteristics as care setting or facility, the type of care or specialty, and the specific segment(s) of the patient record for which the terminology was needed. For each term submitted, participants searched for the single most closely related concept present in the set of existing controlled vocabularies and then indicated whether this concept had exactly the same meaning as their term or was broader, narrower, or had some other relationship. After the initial data collection, a team of subject experts reviewed the test data to determine if participants had missed synonyms or closely related terms (e.g., due to limitations in the test application), or had obviously misinterpreted test instructions. The inter-rater reliability among the reviewers was tested and found to be high.

Results

During a six-month period, 63 participants contributed 41,127 terms, which included 32,679 unique "normalized" strings. In other words, about 20 percent of the terms submitted were identical or very similar in form to other submissions, differing only in case, capitalization, word order, punctuation, etc. For example, "WOUNDS," "Wound," and "wound" counted as one normalized string. Based on the general categorization provided by the testers, there was a good distribution of terms needed in different care settings (inpatient, ambulatory, long term, and home care) and for different specialties and types of care. Eighty percent of the terms submitted were needed for parts of the patient record that describe patient conditions (e.g., diseases, symptoms, signs, test results).

Overall, the Vocabulary Test Set contained the exact meaning of 58 percent or 23,837 of the terms submitted. This figure includes only those cases in which the exact meaning of the term submitted was present in a single entry in one or more of the vocabularies in the test set. (As explained below, additional meanings could be represented by combinations of entries from specific test vocabularies that allow such combinations.) The 23,837 terms for which exact meanings were found represented a maximum of 12,707 discrete concepts, reflecting synonyms, slight variations, and duplicates among the terms submitted. The percentages of exact matches found varied by specialty from a low of 45 percent for veterinary medicine to a high of 71 percent for ophthalmology.

For an additional 41 percent of the terms submitted, testers and reviewers found a closely related concept which was generally narrower in meaning. Many of these differed from an entry in the test set only by a single modifier. Thus, 99 percent of the terms submitted had exact meanings or related concepts in the existing controlled vocabularies included in the test.

Twenty-nine different vocabularies contained some of the exact meanings of terms submitted in the test. No single vocabulary accounted for more than 63 percent of the terms or 57 percent of the unique concepts for which exact matches were found by testers or reviewers. Only SNOMED International and the Read Codes contained the meanings of more than 60 percent of the terms or more than 50 percent of unique concepts for which exact matches were found.

To provide a second level of review, the developers of SNOMED International and the Read Codes were given a random sample of terms for which exact matches had not been found by testers or reviewers. Their task was to determine if additional matches were present as single entries in the most current versions of their vocabularies or if the exact meanings of the terms could be represented by a valid combination of entries in their systems. Based on the results from this sample, it is probable that the percentage of exact matches would have increased from 58 to 65 percent if the most current versions of SNOMED International and the Read Codes had been included in the test set. The percentage of exact meaning matches would increase to 79 percent if valid synonymous combinations of entries from these systems were included.

Conclusions

The results of the LSVT indicate that existing vocabularies contain the bulk of the concepts and terms needed to describe patient problems and conditions. The combination of existing controlled vocabularies has better coverage than any single system. These findings corroborate the results of previous vocabulary studies.[4,5] The organizational and technical success of the test augurs well for the use of the Internet in collaborative vocabulary efforts and in informatics research in general.

Implications

The development of less ambiguous and more comparable patient data will require a target vocabulary or—more feasible and likely—a set of nonoverlapping controlled vocabularies that together cover the concepts needed to document patient problems and the process of care.[8] As use of the UMLS Metathesaurus and lexical resources have illustrated, the ability to use such target vocabularies in operational systems can be improved by rich synonymy obtained by links to other vocabularies and by creative use of lexical tools and techniques. If most of the work of "creating" the controlled vocabulary needed to describe patient problems and conditions has already been accomplished, the pressing priority is to develop a workable strategy that will eliminate unnecessary duplication of effort among vocabulary developers and provide a stable source of support for ongoing development and enhancement of key complementary vocabulary systems. It is also essential to establish a working feedback loop between the clinical information system builders and users on the one hand and vocabulary developers on the other. Such interaction is critical because, to some extent, the structure of computer-based patient records will determine the range of controlled vocabulary that is needed.

As agreement is reached on at least some of the controlled vocabularies that will make up the target set for patient data, work can proceed on developing "standard" mappings and algorithms between these vocabularies and administrative and statistical codes (e.g., ICD-9-CM and CPT). This will allow the administrative codes for disease conditions, procedures, etc. to be generated at least semiautomatically from more detailed electronic patient data captured during the process of diagnosis and treatment. The LSVT results provide some evidence that this fundamental change in the way administrative health data is produced may not be very long in coming.

Notes

1. Unified Medical Language System, UMLS, and Metathesaurus are registered trademarks of the National Library of Medicine.

2. The UMLS Knowledge Sources, lexical programs, and the LSVT test data are available free of charge to those who sign the license agreement for use of the UMLS products. For some uses of some of the individual vocabularies contained in the UMLS Metathesaurus, users will have to enter into separate agreements, which may involve fees, with the producers of the individual vocabularies. Current information about the UMLS project, including a comprehensive bibliography and instructions for obtaining the UMLS resources, is available from NLM's Web site at http://www.nlm.nih.gov.

3. Humphreys, B.L., A.T. McCray, and M.L. Cheh. "Evaluating the Coverage of Controlled Health Data Terminologies: Report on the Results of the NLM/AHCPR Large-Scale Vocabulary Test." *Journal of the American Medical Informatics Association* 4, no. 6 (1997): 484-500.

4. Chute, C.G., S.P. Cohn, K.E. Campbell, D.E. Oliver, and J.R. Campbell for Computer-based Patient Record Institute's Work Group on Codes and Structures. "The Content of Clinical Classifications." *Journal of the American Medical Informatics Association* 3, no. 3 (1996): 224-33.

5. Campbell, J.R., P. Carpenter, C. Sneiderman, S. Cohn, C.G. Chute, and J. Warren for CPRI Work Group on Codes and Structures. "Phase II Evaluation of Clinical Coding Schemes: Completeness, Taxonomy, Mapping, Definitions, and Clarity." *Journal of the American Medical Informatics Association* 4, no. 3 (1997): 238-51.

6. Humphreys, B.L., W.T. Hole, A.T. McCray, and J.M. Fitzmaurice. "Planned NLM/AHCPR Large-Scale Vocabulary Test: Using UMLS Technology to Determine the Extent to Which Controlled Vocabularies Cover Terminology Needed for Health Care and Public Health." *Journal of the American Medical Informatics Association* 3, no. 4 (1996): 281-7.

7. McCray, A.T., M.L. Cheh, A.K. Bangalore, K. Rafei, A.M. Razi, G. Divita, and P.Z. Stavri. "Conducting the NLM/AHCPR Large Scale Vocabulary Test: a Distributed Internet-based Experiment," in *Proceedings of the 1997 AMIA Annual Fall Symposium*, ed D. Masys. Philadelphia, PA: Hanley & Belfus, 1997: pp. 560-564.

8. Board of Directors of the American Medical Informatics Association. "Standards for Medical Identifiers and Messages Needed to Create an Efficient Computer-based Medical Record." *Journal of the American Medical Informatics Association* 1, no. 1

Betsy L. Humphreys *is assistant director for health services research information at the National Library of Medicine, Bethesda, MD.*

Article Citation:

Humphreys, Betsy L. "Do Existing Controlled Vocabularies Contain Terminology Needed for Patient Records?" *Journal of AHIMA* 69, no. 5 (1998): 30-33.
