Using Patient Data to Retrieve Health Knowledge

James J. Cimino, MD;[1] Mark Meyer, MPH, MD;[1] Nam-Ju Lee, MS;[2] Suzanne Bakken, RN, DNSc[1,2]
[1]Department of Biomedical Informatics and [2]School of Nursing
Columbia University, New York, New York, USA

Background: *We sought to study a variety of terminologic approaches to the use of clinical data for searching on-line information resources.*
Methods: *We used a collection of narrative text and coded data to search a variety of text-based, concept-based, and concept-indexed resources.*
Results: *Automated retrievals using original terms produced ample results. Quality of the results varied with the resource. Terminology translations were difficult to accomplish and produced variable results.*
Conclusions: *Current resources support automated retrieval; however, achieving quality results varies with the terms and the resources, with term translation productive only in select situations.*

INTRODUCTION

One of the challenges to the integration of computer systems is the translation of information from the source system into a form is recognized by the target system. This is especially true for infobuttons, which are links between clinical information systems and online knowledge resources that use patient data to retrieve relevant health information.[1] Others have addressed translation by integrating resource links into the data dictionary used by their system[2] and by automated translation of source terms to resource index terms,[3] but the results of retrievals using these methods have not been reported. We have previously described six linkage methods[4] and now describe our experience with term translation and retrieval using those methods.

METHODS

Two of the infobutton integration methods[4] do not require translation because the meaning of the information being transferred is implicit in the integration process.* The approach to translation required to support the remaining methods depends on the kind of patient data being used (coded or uncoded) and the method by which the information in the resource is accessed, indexed, or otherwise organized (narrative text or controlled terminology).

* *Simple links* use a predetermined relationship between a concept and a resource document. *Calculators* obtain data as specific parameters, obviating the need for translation.

Patient Data Sources and Representation

The patient data used in this study included radiology reports, medication orders, laboratory tests, microbiology sensitivity tests, and microbiology results (see Table 1). The radiology reports were represented as uncoded narrative text, while the remaining data were coded with local terminologies that are specific to the pharmacy and laboratory systems. The local terminologies have been incorporated into our Medical Entities Dictionary (MED), which contains additional knowledge, such as medication ingredients and test analytes.[5]

Health Knowledge Resources

Retrievals were applied against a variety of available knowledge resources (see Appendix). Most (PubMed, Rx List, Up to Date, Micromedex, Labtests Online, the National Guidelines Clearinghouse (NGC) and One Look) were accessed using a text-based *simple search* method[4] included in their user interfaces. Two resources (CPMC Lab Manual and Lexicomp) are sets of documents, with a particular concept (medication or laboratory test) for each document; we refer to these as *concept-based resources*.[4] One source (PubMed) provides *concept-based searching* using a controlled terminology.[4]

Uncoded Data and Narrative Text Resources

We used MedLEE, a medical language processor,[6] to identify findings and diagnoses in radiology reports.[7] Performing simple searches with these terms involved inserting them into appropriate links. To search the One Look dictionary for "infiltrate", for example, one uses the link:

http://www.onelook.com/?w=infiltrate

We used MedLEE to obtain all findings and diagnoses from a collection of anonymized reports, processing reports until we obtained 100 unique findings/diagnoses. We then used these terms to search against four simple-search resources (PubMed, Up to Date, National Guidelines Clearinghouse (NGC), and OneLook Dictionary).

Coded Data and Narrative Text Resources

Searching text resources with coded data is not always as simple as using the term name as a search term. As Table 1 shows, some term names are verbose (such as "UD AMIKACIN 1 GM VIAL")

Data Source	Source Terminology	# Terms	Example Term
Radiology Reports	Narrative Text	N/A	"…**infiltrate** is seen in the **left upper lobe**."
Medications	Local Formulary	15,311	UD AMIKACIN 1 GM VIAL
Laboratory Tests	Local Laboratory	6,133	AMIKACIN, PEAK LEVEL
Sensitivity Tests	Local Laboratory	476	AMI 6 MCG/ML
Microbiology Results	Local Laboratory	2,173	ESCHERECHIA COLI

Table 1: Representation of Patient Data Used for Retrieval (UD stands for "Unit Dose")

while others are cryptic ("AMI 6 MCG/ML"). We generally found the microbiology term names to be appropriate for searching, but used other attributes for the other coded data: analyte names for laboratory test terms, ingredients for medication terms, and antibiotic names for antibiotic sensitivity test terms.

We used appropriate names (as above) for 100 searches against each text resource: Lab Tests Online for test terms (using analyte names), Rx List and Micromedex for medications and sensitivity terms (using ingredient names), and PubMed and Up to Date for microbiology results (using the names of the results themselves). Most searches returned sets of citations (PubMed), documents (Lab Tests Online) or topics (Up to Date). We counted the results in each set and carried out a manual qualitative assessment.

The two medication resources, Rx List and Micromedex, produced structured displays of heterogeneous links, rather than lists of citations or documents. This format prevented us from simply counting the results, since many links were ancillary to the topic of interest or were redundant. Instead, we determined whether any of the links were appropriate (true positives). Where no true positive links were found, we manually searched the resource using synonyms (such as known brand names); if we obtained true positive results using this method, we considered the original searches to be false negatives.

Coded Data and Controlled Term Resources

We studied four methods for using coded patient data to retrieve information from concept-based or concept-searchable resources.

Manual mapping to controlled terms: As part of a previous effort,[5] microbiology result terms were manually mapped to the National Library of Medicine's (NLM's) Medical Subject Headings (MeSH).[8] These mappings were stored in our local dictionary (MED) and used to automatically translate microbiology terms to MeSH.

Automated translation using the UMLS: As part of a previous effort,[5] the analyte terms used to describe laboratory test terms were manually mapped to terms in the Logical Objects, Identifiers, Names and Codes (LOINC).[9] The LOINC terms are included in the Metathesaurus of the NLM's Unified Medical Language System (UMLS),[10] with each term being mapped to a Concept Unique Identifier (CUI) corresponding to its meaning. Translation to a target terminology can be accomplished by identifying target terms that are mapped to the same CUI. We exploited this ability to translate the LOINC analyte terms to MeSH terms.

Automated mapping to a concept-based resource: Lexicomp is a drug information resource consisting of a set of discrete hyperdocuments, each of which corresponds to a single drug. Documents include Patient Advisory Leaflets (PALs) for adults, PALs for children, and drug images. We used an automated method (with manual review) to map our local formulary medication terms to the medications described in the documents. First, drug names, trade names and synonyms were extracted from the Lexicomp drug information pages using a Perl script. These terms were automatically matched against our medication names and the matches were then manually reviewed by two physicians (JJC and MM) to determine their veracity. Once a match was verified, the document identifier (in this case, the Lexicomp document name) was added to the MED as the Lexicomp translation. Because the MED is a hierarchy, those medication terms that did not match Lexicomp documents directly were able to inherit translation assignments from their ancestors. Medication terms that were assigned identifiers for multiple Lexicomp documents in the same class (Adult PALs, Pediatric PALs, or Drug Images), either through direct matching, inheritance, or a combination of these, were manually reviewed to resolve conflicts.

Manual mapping to a concept-based resource: Like Lexicomp, the CPMC Lab Manual consists of a set of hyperdocuments, each corresponding to a particular concept – in this case, laboratory tests. Although both the Manual and the laboratory test terms originate from the same hospital department, there is no coordination between them. Therefore, one nurse (NJL) and one physician (JJC) mapped terms to Manual documents manually; occasional discrepancies were resolved through mutual discussion. Where possible, we mapped the MED's test classes, since successful matches allowed us to

assign matches to the individual test terms in the classes. Matches were stored in the MED as attributes of the test terms.

For each of the above methods, we determined the percent of source terms that were successfully translated and then examined the usefulness of the translation for retrieval. For each of the four terms set, we selected 100 unique, successfully-translated terms and performed retrievals against the relevant resources (PubMed for microbiology terms, PubMed for laboratory test terms, Lexicomp Adult PALs for medication terms, and the CPMC Laboratory Manual for laboratory test terms) and examined the results.

RESULTS

The results of the translation and searches are summarized in Table 2. The links used to carry out each of the 1,588 searches, along with the assessment of each result, can be found in the Appendix.

Uncoded Data and Narrative Text Resources

We processed 20 reports, containing an average of 6.1 unique diagnosis/finding terms (range: 1 to 16 unique terms per report); together, these 20 reports contained 100 unique terms. As shown in Table 2, all four of the relevant resources were able to produce some results most or all of the time.

In general, PubMed and NGC produced large numbers of results, with all results containing the term of interest but not necessarily discussing it. For these resources, relevance improved as the term became more specific. Thus terms like "compression fractures" tended to yield smaller numbers of results (4,269 PubMed citations and 33 NGC guidelines) that appeared relevant, while nonspecific findings such as "abnormality" produced large numbers of relatively uninteresting results (330,855 citations and 493 guidelines). Meanwhile, the more topic-oriented resources, Up to Date and One Look, tended to produce small numbers of highly relevant results.

Coded Data and Narrative Text Resources

We were able to randomly select 100 terms each for medications, microbiology results, and laboratory tests. However, although there are 476 sensitivity tests, they reference only 94 unique antibiotics, so that a total of 394 searches were performed using coded data against narrative sources.

In general, searches with drug ingredient names yielded high-quality results. Rx List produced long lists of odd false-positive matches, but true positives were found for most terms (88% of medications and 79% of antibiotics); in some cases (4% of medications and 6% of antibiotics), the original term failed to produce a true positive, but the term was found through manual search with a synonym. Micromedex produced much more focused results, with true positives found for most terms (89% of medications and 96% of antibiotics); additional terms were found (6% of medications and 1% of antibiotics) through manual search using synonyms. Of the 36 terms that were missed by one or both resources (true and false negatives), 9 were missed by both, 22 were missed by Rx List only and 5 were

Terms from Data Source	Searches Performed	Retrieval Success
100 Findings and Diagnoses from 20 Radiology Reports	100 PubMed	100 % (92,440)
	100 Up to Date	82% (28.6)
	100 NGC	95% (119)
	100 One Look	81% (25.8)
100 Medication Terms (using ingredient names)	100 Rx List	95% [.88/.04]
	100 Micromedex	100% [.89/.06]
94 Sensitivity Test Terms (using antibiotic names)	94 Rx List	85%[.79/.06]
	94 Micromedex	97% [.96/.01]
100 Microbiology Result Terms	100 Up to Date	94% (1.4)
	100 PubMed	100% (3,328)
	100 PubMed (using MeSH translation)	100% (18,036)
100 Lab Test Terms (using analyte names)	100 Lab Tests Online	73% (133)
	100 PubMed	99% (84,633)
	100 PubMed (using MeSH translation)	100% (90,656)
100 Medication Terms	100 Lexicomp (using document identifiers)	96% (1)
100 Laboratory Test Terms	100 Lab Manual (using document identifiers)	94% (1)

Table 2: Summary of results of using patient data to search information resources. Target resources and terminologies are identified in the Appendix (NGC is National Guidelines Clearinghouse). Retrieval success is represented as percent of terms that successfully retrieved any results; numbers in parentheses indicate average numbers of results (citations, documents, topics, definitions, etc., depending on the target resource) for those searches that retrieved at least one result. Results for Rx List and Micromedex are difficult to quantify, because they provided heterogeneous lists of links; rather than provide link counts, we assessed the true positive and false negative rates, shown in brackets.

missed by Micromedex only.

Searches of the other focused-content resources, Up to Date (using microbiology result terms) and Lab Tests Online (using test analyte names), produced positive results 94% and 73% of the time, respectively. The topics returned by Up to Date for each search were much fewer than the searches with radiology report terms: 1.4 vs. 28.6, on average. Lab Tests Online's result sets tended to be larger but were sorted in order of decreasing relevance.

As with the searches of radiology report terms, the searches of PubMed using coded terms produced large result sets. However, the microbiology result terms produced sets that were significantly smaller than the sets produced with radiology report terms and laboratory test analytes.

Coded Data and Controlled Term Resources

Manual mapping to controlled terms: We were able to identify MeSH terms for 1,028 (47.3%) of the microbiology result terms. The results of searching PubMed with a random sample of 100 of these terms produced result sets that were approximately 6 times larger than those produced with the original terms. Closer inspection shows the differences to be even more striking: after excluding the 33 cases where the original terms and the MeSH translations were the same (which obviously yielded identical search results), there were 7 cases where the original term produced slightly larger sets than the MeSH translation (6,007 vs. 4,159, on average) and 60 cases where the MeSH translation produced larger sets than the original terms (25,585 vs. 855, on average; a 30-fold increase). In general, though, these larger search results may not represent improved recall: most of the cases involved translation of a genus-species name to a MeSH term that was genus-only. For example, searching PubMed with the microbiology result term "CANDIDA VISWANATHII" retrieved 28 citations, while searching with its MeSH translation ("Candida") retrieved 33,440 citations. Depending on the question being asked, these additional citations may or may not be relevant.

Automated translation using the UMLS: We were able to identify LOINC terms for 940 (90.3%) of the analytes used to describe our laboratory tests. Using the UMLS, we were able to automatically map 485 (51.6%) of these to MeSH terms. The results of searching PubMed with a random sample of 100 of these MeSH terms produced sets that, on average, were about the same as searching with the original terms (90,656 vs. 84,633). However, this was due to the fact that in 72 of the cases the MeSH translation was identical to the original term. When the terms were different, the original terms produced slightly larger sets in 16 cases (225,453 vs. 215,216), and the MeSH terms produced much larger sets in 12 cases (101,992 vs. 31,101).

Automated mapping to a concept-based resource: Initially, the matching algorithm suggested 22,694 matches. After removal of 1,610 matches that were judged to be erroneous, a total of 21,084 matches for medication terms were found: 8,888 for Adult PALs, 3,329 for Pediatric PALs, and 8,867 for Drug Images; 10,422 (68.1%) of medications had at least one match to a Lexicomp document. When we searched Lexicomp Adult PALs with a random sample of 100 terms, we found that in general there was one appropriate document per term; however, in 4 cases, we found that the link contained in the MED pointed to an incorrect or missing document.

Manual mapping to a concept-based resource: Manual matching of laboratory test terms against the lab manual identified at least one match for 4,282 (69.8%) terms. In many cases, multiple matches for individual tests were found; for example, "Amniotic Fluid Culture" matched documents for "Body Fluid Culture, Routine, Aerobic/Anaerobic", "Body Fluid Culture, Mycobacteria", "TB Culture, Body Fluid", "Body Fluid Culture, Fungus", and "Body Fluids Culture". We reviewed these multiple matches and manually selected one that appeared to be the most appropriate. When we searched the CPMC Lab Manual with a random sample of 100 terms, we found that in general there was one appropriate document per term; however, in 6 cases, we found that the link contained in the MED pointed to an incorrect or missing document.

DISCUSSION

Effective integration of information systems requires careful attention to the means by which data are translated from the source system into a form recognized and understood by the target system. The purpose of this study was to examine a variety of methods for using patient data from electronic health records as search parameters to on-line health knowledge resources. We used a set of 694 clinical data items to search nine resources. In some cases, we carried out translations using a mixture of manual, knowledge-based, and UMLS-based techniques. In all, we carried out 1,588 searches. We made qualitative observations about the search results; measurements of precision and recall, however, were not relevant, since we were studying only the technical aspects of automated retrievals,

rather than attempting to answer specific clinical questions.

In general, the searches were technically successful – 1,453 (91.5%) returned results that contained true positive results. However, of the 135 retrievals that produced no positive results, at least 41 of these were false negatives. Thus, effective automated retrievals remain difficult to achieve, even with controlled terminologies.

In most cases, we used text terms and names of terms from local terminologies as our search terms; in other cases we used formal knowledge about the terms (test analytes, medication ingredients and antibiotics used in sensitivity tests) to provide more appropriate names. These strategies generally worked well. Manual and automated (UMLS-based) translation to a standard terminology (MeSH) did not appreciably improve the retrievals and, in some cases, made them worse. Manual and automated mapping from local terms to concept-specific resources (Lexicomp and the CPMC Lab Manual) facilitated effective retrievals, although mapping errors occasionally occurred.

Our experience shows that, at least for automated retrievals of health information using patient data, the "terminology problem" remains significant. Data are in nonstandard forms, and translation is problematic. Translation may not be advantageous because most sources don't have a target terminology and, when they do, the translation may make things worse rather than better.

As health information systems begin to embrace widely available controlled terminologies and information resources begin to make more use of controlled terms for indexing their material and performing searches, we believe that approaches such as ours to automated patient-data-based knowledge retrieval will be widely applicable. Furthermore, the recent growth in the UMLS is supporting improved translation rates, when compared with previous attempts.[11]

This project attempted to answer the technical question of whether patient data could be used for automated knowledge retrieval. Further work is needed to determine how terms from the patient record can be used to enhance searching and how to address precision and recall for specific user questions; we believe our collection of methods and tests provides some basis for carrying out this work.

CONCLUSIONS

Current resources support automated retrieval; however, achieving quality results varies with the terms and the resources, with term translation advantageous only in select situations.

Appendix

Tables describing the resources used in this study, and the detailed results of all 1,588 searches performed (including links that execute the actual searches) are included on the Web at:

www.dbmi.columbia.edu/cimino/2005amia-data.html

Acknowledgments

This work is supported in part by NLM grants R01LM07593 and R01LM07659 and NLM Training Grants LM07079-11 and P20NR007799. The authors thank Andria Cimino for editorial assistance.

References

1. Cimino JJ, Elhanan G, Zeng Q. Supporting Infobuttons with Terminological Knowledge. *JAMIA*; 1997;4 (Suppl):528-532.
2. Ruan W, Burkle T, Dudeck J. An object-oriented design for automated navigation of semantic networks inside a me. dical data dictionary. *Artificial Intelligence in Medicine*. 2000; 18(1):83-103.
3. Reichert JC, Glasgow M, Narus SP, Clayton PD. Using LOINC to link an EMR to the pertinent paragraph in a structured reference knowledge base. *Proc AMIA Symp*. 2002:652-6.
4. Cimino JJ, Li J, Allen M, Currie LM, et al.. Practical Considerations for Exploiting the World Wide Web to Create Infobuttons. *Proc of Medinfo 2004*:277-281
5. Cimino JJ. From data to knowledge through concept-oriented terminologies: experience with the Medical Entities Dictionary. *JAMIA*; 2000;7(3):288-297.
6. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. *J Am Med Inform Assoc*. 2004 Sep-Oct;11(5):392-402.
7. Janetzki V, Allen M, Cimino JJ. Using natural language processing to link from medical text to on-line information resources. *Proceedings of Medinfo 2004*:1665.
8. www.nlm.nih.gov/pubs/factsheets/mesh.html
9. Forrey AW, McDonald CJ, DeMoor G, et al. Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results. *Clin Chem*. 1996 Jan;42(1):81-90.
10. www.nlm.nih.gov/research/umls
11. Cimino JJ, Johnson SB, Peng P, Aguirre A: From ICD9-CM to MeSH using the UMLS: A How-to Guide. In Safran, C, ed.: *Proceedings of the SCAMC*. 1993:730-734.