Automated Knowledge Extraction from the UMLS

Qing Zeng M.S., James J. Cimino, M.D.
Department of Medical Informatics
Columbia University, New York, New York

This paper presents our work in extracting disease-chemical relationship knowledge from the UMLS Co-occurrence table (MRCOC) using an automated method. We evaluated the quality of the knowledge from UMLS MRCOC by comparing it with knowledge from other sources: For disease-lab chemical relationships, knowledge was obtained from a decision support system (DXplain) and our own knowledge base of medical terminology (MED) through automated processes. For disease-drug chemical relationships, knowledge was manually acquired from the medical literature. Evaluations showed that the UMLS MRCOC knowledge has good sensitivity, especially regarding disease-drug relationships. We are using this knowledge to produce disease-specific views of patients' electronic patient record.

INTRODUCTION

Most current Clinical Information Systems (CISs) present results and orders organized according to ancillary departments and chronological order. These department and time-oriented views do not necessarily support information retrieval for every kind of clinical task. Other views can be envisioned which center around clinical concepts such as diagnostic strategies, therapeutic goals, etc. We refer to these as *concept-oriented views*. Examples of such views include the problem-oriented views[1], prototypical question-oriented views[2] and anticipatory patient data displays[3].

Traditionally, concept-oriented views are manually created and maintained. However, efforts have been made to automate the view generation proccess[2]. At Columbia-Presbyterian Medical Center (CPMC), we have been working on a system that generates concept-oriented views of our electronic medical record (EMR) system in a dynamic way.

Generating these views requires our system to have knowledge about relationships between concepts used in the EMR. For example, in order to determine if cardiac enzyme tests should be included in a view of heart disease, the system needs to know that cardiac enzymes are related to heart diseases and which laboratory tests measure those enzymes.

For disease-oriented views, the knowledge of the relationships between diseases and chemicals are important because chemicals act as pharmaceutical components of drugs as well as analytes of laboratory tests. Such knowledge supports the views that link relevant diagnostic laboratory tests and drug prescriptions with the diseases.

Acquiring the disease-chemical relationship knowledge is challenging because of the amount of knowledge required and the need for constant update due to new developments in medical research. One approach we are testing is to extract the knowledge from existing resources using automated methods.

Previous attempts have been made to use the co-occurrence of keywords in MEDLINE[4] citations as evidence of inter-concept relationships[5]. Such methods, which depend on executing searches and analyzing their results, are laborious and time-consuming. Fortunately, the Unified Medical Language System (UMLS)[6] now provides this co-occurrence information in its MRCOC file[6]. To assess the quality of the UMLS as a source of this kind of knowledge, we selected another source - a diagnostic expert system DXplain[7] for comparison. Since DXplain does not contain drug information for diseases, we also manually collected knowledge in that domain from the medical literature (textbooks and review articles). In this paper, we describe the knowledge extraction methods and the evaluation of knowledge acquired from UMLS by comparing it with the knowledge from other sources.

METHODS

Knowledge Extraction from UMLS

The UMLS contains information about the co-occurrence of "concepts that were designated as principal or main points in the same journal article"[6] from MEDLINE. This information is stored in the MRCOC table which is a publicly available text file. The following are a few sample entries from the MRCOC table:

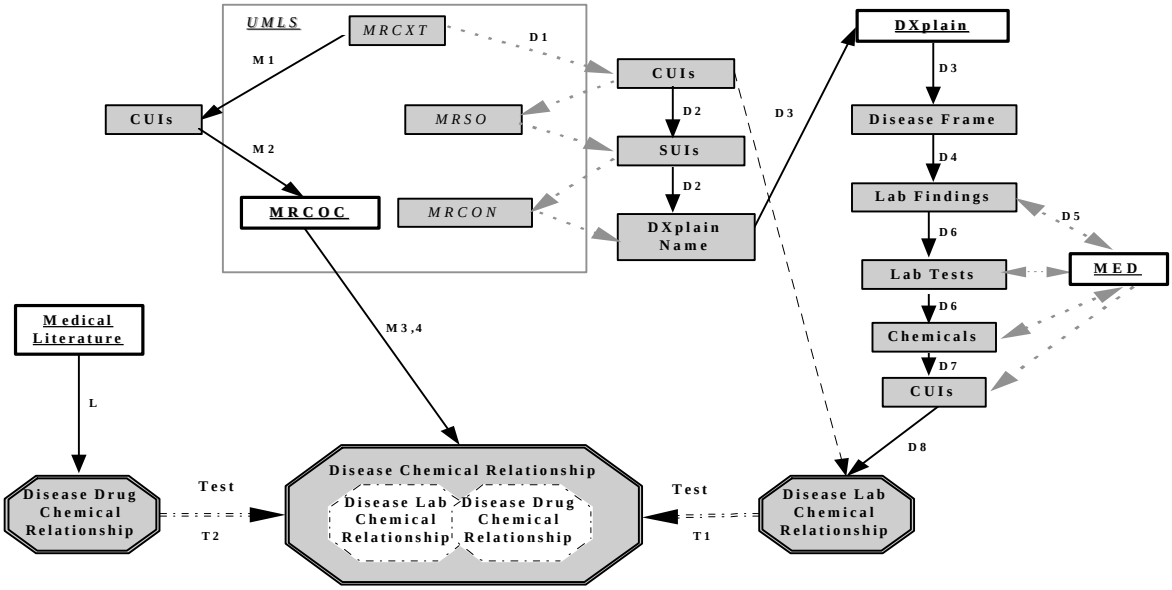

Figure 1 This figure shows the methods of knowledge extraction from UMLS, DXplain and medical literature. It also shows that the knowledge from DXplain and Medical Literature are used to test knowledge from UMLS MRCOC. (Numbers refer to steps documented in the text.)

C0000165|C0032460|MBD|L|4|DF=3,ME=1|
C0000165|C0032460|MED|L|1|BL=1,ME=1|
C0032460|C0000165|MBD|L|4|ME=2,EN=1,...,PP=1|
C0032460|C0000165|MED|L|1|BL=1,DT=1|

The first and second fields are the UMLS concept unique identifiers (CUIs) of two concepts that co-occurred. The third field indicates the source database (MBD: MEDLINE 1985-91, MED: MEDLINE 1992-95). The fourth field is the type of co-occurrence. (L: co-occurrence of primary or main subject headings) The fifth field contains the number of co-occurrences. The last field contains MESH subheadings[8] that belong to the first concept in each entry and are therefore different for each direction of the relationship. Note that MRCOC (as shown in this sample) contains reciprocal entries for each entry from MEDLINE.

In order to extract potential disease-chemical relationships from the UMLS, we used the following procedures:

1. Identify the disease and chemical concepts using MESH hierarchy. This hierarchy is represented in MESH tree addresses that can be found in the UMLS file MRCXT[6]. We collected the CUIs corresponding to the MESH identifiers of diseases and chemicals. [M1-Figure 1]

2. Select entries from MRCOC that contain one disease CUI and one chemical CUI. [M2-Figure 1]

3. Identify MESH subheadings that could imply "disease-analyte" or "disease-drug" relationship, for example, *diagnostic use* and *adverse effects respectively*. Select only entries that contain such subheadings. [M3, 4-Figure 1]

4. Merge entries regarding the same pair of disease-chemical into a single entry. As seen in the sample entries from MRCOC, information about one pair of disease and chemical is stored in different entries based on the direction of relationship and the source. (In this study, we combine co-occurrences across the two MEDLINE databases in MRCOC and of reciprocal entries.) [M3, 4-Figure 1]

Knowledge Extraction from DXplain

DXplain is a diagnostic assistant system which stores disease related clinical findings (including test results) in disease descriptions. To extract disease-analyte (chemical) relationships from DXplain, we used knowledge from the Medical Entity Dictionary[9] (MED), a semantic network as described below.

We used the following procedure to extract potential disease-chemical relationships from DXplain:

1. Select CUIs of interest. [D1-Figure 1]

2. Find the string unique identifiers (SUIs) of DXplain names for the CUIs in UMLS MRSO[6] and

translate the SUIs to DXplain names using UMLS MRCON[6]. [D2-Figure 1]

3. Obtain disease descriptions of the concepts from DXplain[7]. [D3-Figure 1]

4. Attain lab findings and the frequencies with which they are associated with a disease from DXplain disease descriptions. The following is an excerpt from a disease description: [D4-Figure 1]

NEPHROPATHY, OBSTRUCTIVE
--------- LABORATORY FINDINGS --------
RARELY: blood urea nitrogen elevated; creatinine, elevated.

5. Translate the lab findings to MED concepts. [D5-Figure 1]

6. Retrieve the lab tests that define the findings and identify the analytes of the tests. This is done by following the semantic links in the MED. [D6-Figure 1]

7. Translate the MED analytes to UMLS concepts. [D7-Figure 1]

8. Create a table of the DXplain disease CUIs, the related analytes CUIs, and the frequency of the lab findings' association with the disease. [D8-Figure 1]

Knowledge Extraction from Literature

For the diseases that were randomly selected to test MRCOC disease-chemical relationships, we searched for related drugs in textbooks and review articles. The drugs were mapped to UMLS concepts to construct a list of disease-chemical relationships. The process was done manually. [L-Figure 1]

Evaluation

We used the disease-chemical knowledge acquired from DXplain to test the sensitivity of MRCOC knowledge regarding disease-lab chemical relationships. Among the diseases that both DXplain and MRCOC cover, ten diseases were randomly selected. For each disease, entries of disease-chemical relationship were selected from the previously extracted MRCOC and DXplain knowledge. The percentage of DXplain disease-chemical entries also existing in MRCOC (i.e. sensitivity) was calculated for the random sample. [T1-Figure 1]

We then used the disease-chemical knowledge collected from textbook and review articles to test the sensitivity of MRCOC knowledge regarding disease-drug chemical relationship. Six diseases were randomly selected from the diseases MRCOC covers. Disease-chemical relationships for these diseases were constructed from medical literature and selected from the previously extracted MRCOC knowledge. Sensitivity was calculated for the random sample. [T2-Figure 1]

When comparing knowledge extracted from different sources, granularity of chemical terms is a problem. For instance, one source may link angina with the drug class beta-blocker and not mention the specific drugs in that class. Another may list individual drugs without mentioning the class. As a solution, we used the *explode* strategy: Descendants of a chemical in the MESH tree were linked to the diseases to which the chemical ancestor is linked. The descendants also inherited the co-occurrences of their ancestors. As an example, if we explode the angina-beta-blocker relationship, all descendents of beta-blocker would be linked to angina and relationships such as angina-propranolol would be created.

Three sensitivity tests for disease-lab chemical relationships were conducted between DXplain knowledge and the MRCOC: unexploded DXplain knowledge vs. the unexploded MRCOC, unexploded DXplain vs. the exploded MRCOC, and exploded DXplain vs. the exploded MRCOC. Three corresponding sensitivity tests for disease-drug chemical relationships were conducted between literature knowledge and the MRCOC. In the cases using exploded DXplain knowledge or exploded literature knowledge, a relationship is considered matched if a descendent of the relationship was found in MRCOC. Using the beta blocker propranolol as an example, if angina-beta blocker is in our literature knowledge and angina-propranolol is in the MRCOC, we would consider angina-beta blocker a match.

We theorized that the distribution of the number of co-occurrences in the MRCOC entries which were matched with DXplain or literature knowledge differ from the distribution in the random samples and analyzed the distributions.

According to DXplain, there are findings that are usually associated with a disease and there are ones that are rare. We also theorized that the MRCOC would be more sensitive to common findings.

RESULTS

UMLS Knowledge Extraction

3597 disease and 6126 chemical concepts were identified in the 1997 MESH. 994758 lines in MRCOC contain one disease and one chemical. After removing the lines with irrelevant subheadings, 782906 lines remained. After merging co-occurrences of the same disease-chemical relationship, 389655 relationships were extracted.

3388 diseases and 5683 chemicals were linked through these relationships.

DXplain Knowledge Extraction

We chose to focus on cardiovascular and urologic diseases. 182 DXplain disease descriptions were obtained. Following the links from lab findings in the disease descriptions to lab tests and then to analytes, 336 pairs of disease-analyte relationships were extracted.

Evaluation

To test MRCOC's sensitivity regarding disease-lab chemical knowledge, 54 relationships were extracted from DXplain and 2820 relationships were extracted from MRCOC for the random sample of ten diseases. Table 1 shows sensitivity values. Figure 2 illustrates the distribution of co-occurrences of in MRCOC sample, in MRCOC sample matched with DXplain and in exploded MRCOC sample matched with DXplain.

Table 1: The MRCOC sensitivity regarding lab chemical for 10 random diseases.

	Found (Percentage)
DXplain vs. MRCOC	31 (57.3%)
DXplain vs. Exp. MRCOC	36 (66.7%)
Exp. DXplain vs. Exp. MRCOC	37 (68.5%)

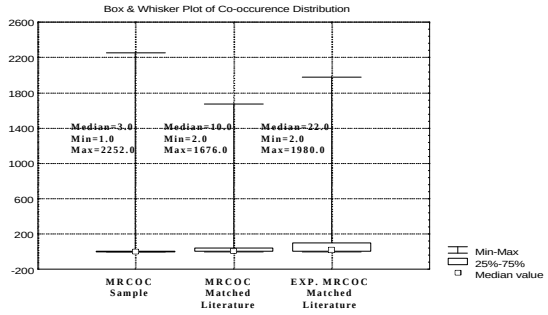

Figure 2. Co-occurrence distributions of entries in MRCOC sample, MRCOC entries matched with literature, exploded MRCOC entries matched with literature

Our observation showed that MRCOC is not more sensitive to findings that are usually related to a disease.

To test of MRCOC's sensitivity regarding disease-drug chemical knowledge, 43 relationships are extracted from medical literature and 1855 relationships are extracted from MRCOC for the random six diseases. Table 2 shows sensitivity values. Figure 3 illustrates the distribution of co-occurrence of related concepts in MRCOC sample, in MRCOC sample matched with medical literature and in exploded MRCOC sample matched with medical literature.

Table 2: The MRCOC sensitivity regarding drug chemical for 10 random diseases.

	Found (Percentage)
Literature vs. MRCOC	35 (81.4%)
Literature vs. Exp. MRCOC	37 (86.0%)
Exp. literature vs. Exp. MRCOC	40 (93.0%)

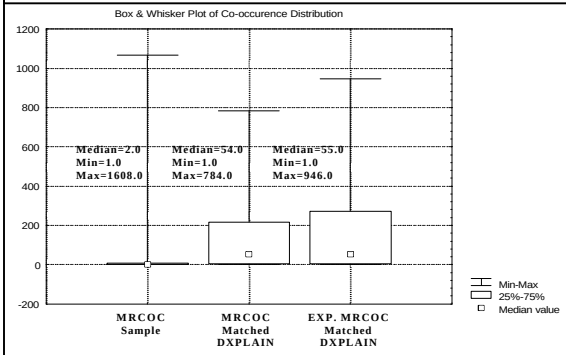

Figure 3. Co-occurrence distributions of entries in MRCOC sample, MRCOC entries matched with DXplain.

Both Figure 2 and 3 show that the matched disease-chemical relationships tend to have a high number of co-occurrences than the random sample.

DISCUSSION

We were able to acquire knowledge of 389655 disease-chemical relationships from MRCOC in UMLS and 336 disease chemical relationships from DXplain using automated methods. Compared to constructing a similar knowledge base manually, this approach is much easier and faster.

There is no gold standard for the sensitivity and specificity of disease-chemical relationship knowledge since relevance can only be defined in context. In this case, the real evaluation would be the evaluation of the performance of our system when the knowledge is put in use. Nonetheless, we could get a general estimation of the sensitivity by comparing knowledge extracted from other sources (expert system and medical literature) with MRCOC knowledge. Specificity is much harder to measure

since the absence of a disease-chemical relationship from one source does not mean it does not exist. For example, a textbook may not mention a drug as treatment for a particular disease while some articles discuss it.

The sensitivity of the disease-drug knowledge from MRCOC appears to be very good (93%). We examined the unmatched disease-drug relationships and found that they all came from two diseases which are not primarily treated by drugs and for which existing drug treatments are sometimes controversial. If we exclude those cases, the sensitivity is nearly perfect.

The sensitivity of MRCOC disease-lab chemical relationship knowledge is not as good (68%). We did observe important disease-lab chemicals missing from MRCOC. A possible reason is that the relationship between diseases and diagnostic lab tests can be expressed in terms of analytes of the tests, tests themselves and lab findings derived from the tests. Unfortunately, lab tests and findings are not indexed in MESH and thus not represented in MRCOC table.

Although we are not able to report specificity at this stage, the experimental views our system generated using MRCOC knowledge suggested poor specificity. Yet, it is important to note that knowledge extracted from MRCOC does not link each disease to every possible chemical. It covers 2% of all possible relationships. Also the sensitivity values prove that these relationships from MRCOC are not random.

It is natural to use the number of co-occurrences to as a filter because known disease-chemical relationships do generally have a higher number of co-occurrences (as observed in Figure 2, 3). The problem is that simply filtering out relationships between concepts that co-occurred just once or twice can reduce sensitivities by over 25%. Further analysis of the co-occurrence and sub-heading distribution is needed to improve specificity.

We are experimenting with using knowledge extracted from MRCOC and DXplain in our system to generate concept-oriented views. Although there is noise in the knowledge, our preliminary experience in using this knowledge to identify lab tests and drug orders related to a disease and create disease specific views is very promising.

MRCOC knowledge is not collected and modeled for view generation. Although this made it inherently challenging to reuse, we explored the possibility because of its broad coverage, free availability and simple electronic format. There are commercial sources that provide knowledge such as drug-disease relationship knowledge that may be more accurate and could also be exploited.

CONCLUSION

We are able to extract disease chemical relationships from UMLS MRCOC and DXplain using automated methods. We evaluated the knowledge from MRCOC using DXplain knowledge and manually extracted knowledge from medical literature. In evaluation, MRCOC showed promising sensitivities (93% in disease-drug chemical relationships and 68% in disease-lab chemical relationships) and the potential to be used for our conceptual view systems.

ACKNOWLEDGMENTS

This work was supported by UMLS contracts from the National Library of Medicine (NLM 95-053/VMS, 001 LM05857-01). The authors thank Dr. Octo Barnett for his generous help in providing DXplain knowledge and Dr. Gai Elhanan for his suggestions.

References

1. Weed LL: Medical records that guide and teach. *N Engl J Med* 1968 Mar 21;278(12):652-657
2. Tang PC, Annevelink J, Suermondt HJ, Young CY: Semantic integration of information in a physician's workstation. *Int J Biomed Comput* 1994 Feb;35(1):47-60
3. Elson RB, Connelly DP: The impact of anticipatory patient data displays on physician decisionmaking: a pilot study. *Proc AMIA Annu Fall Symp* 1997;:233-237
4. http://www.nlm.nih.gov/databases/medline.html
5. Cimino JJ, Barnett GO: Automatic Knowledge Acquisition from MEDLINE. *Methods of Information in Medicine*, 1993; 32(2): 120-130.
6. National Library of Medicine. *UMLS Knowledge Sources - 8th Experimental Edition Documentation.* Bethesda, Maryland: The Library. 1997 Jan.
7. Barnett GO, Cimino JJ, Hupp JA, Hoffer EP: DXplain, An evolving diagnostic decision-support system. *Journal of the American Medical Association*; July 3, 1987; 258(1):67-74.
8. http://www.nlm.nih.gov/mesh/
9. Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of a large controlled medical terminology. *Journal of the American Medical Association*;1994;1(1):35-50.