Automated Knowledge Extraction from MEDLINE Citations

Eneida A. Mendonça, M.D., James J. Cimino, M.D.
Department of Medical Informatics, Columbia University, New York, NY, USA

As part of preliminary studies for the development of a digital library, we have studied the possibility of using the co-occurrence of MeSH terms in MEDLINE citations associated with the search strategies optimal for evidence-based medicine to automate construction of a knowledge base. We use the UMLS semantic types in order to analyze search results to determine which semantic types are most relevant for different types of questions (etiology, diagnosis, therapy, and prognosis). The automated process generated a large amount of information. Seven to eight percent of the semantic pairs generated in each clinical task group co-occur significantly more often than can be accounted for by chance. A pilot study showed good specificity and sensitivity for the intended purposes of this project in all groups.

INTRODUCTION

Previous studies have shown the need of health care providers and patients for access to information pertinent to clinical practice and health–related issues.[1-3] Given the explosion of medical knowledge and faced with the enormous quantity of biomedical literature published annually, physicians find it difficult to keep up-to-date with the advances in medical science.[4] They face a difficulty task of filtering large amounts of information and incorporating evidence to make safe and accurate diagnostic, therapeutic and management decisions. The development of evidence-based decision support tools designed to provide relevant and up-to-date evidence to clinicians has been proposed as a solution to this problem.[5]

There is a need for tools that can facilitate the access to large amounts of information and provide appropriate interactivity. The effective use of technology can be an important facilitator of quality, and utility, in reviewing medical information on the Internet.[6] We believe that the development of personalized access to a distributed digital library can facilitate this process. One challenge in building such a system is the construction of a medical knowledge base to support the search of online medical literature according to individual needs. Such a task can be arduous, in part because of the extensive reviews of medical literature required.[7,8]

Previous research studies have introduced approaches to facilitate knowledge extraction. Some of these studies include automatic extraction from MEDLINE[8] and the UMLS[9]. The method described by Cimino and Barnett depended on executing searches and analyzing their results, and was a laborious and time-consuming task. Zeng and Cimino carried out an automated disease-chemical knowledge extraction based on the co-occurrence of concepts that were designated as principal or main points in the same journal article; this information was provided by the UMLS MRCOC table.[10] The results were promising showing a high estimated sensitivity (93%). Specificity was not estimated.

Pao[11] describes four stages a user goes through before searching for information. First is the recognition of an information deficiency (information problem). Once a problem is identified, an information need should be determined (what is needed to solve the problem). Third is question formulation. Fourth is the conversion of a question into a request. Depending on the results, the user can return to previous stages if necessary. The process described above is analogous to the first step in the practice of evidence-based medicine.[12]

Evidence–based medicine (EBM) focuses on questions related to the central tasks of clinical work: diagnosis, etiology, prognosis, therapy, and other clinical and health care issues. EBM requires the ability to access, summarize, and apply information from the literature to day-to-day clinical problems.[13,14] The first step in this process is to convert information needs into focused questions, formulating a "well-built clinical question".[12] This involves identifying a question that is important to the patient's well-being, is interesting to the physician or health care provider, and that he/she is likely to encounter on a regular basis in his/her practice. According to Sackett, a well-built question usually contain 4 elements: a) a patient or problem being addressed, b) an intervention c) a comparison interventions (optional), and d) an outcome of interest. A fifth element, the type of clinical work (or where clinical questions arise from) is also important in the process of information retrieval. This information is helpful when combining a content search with a methodological quality search,

1067—5027/00/$5.00

which intends to limit the number of studies to those that are most likely to be methodologically sound.[15,16] In this context, Haynes et al.[16] developed optimal MEDLINE search strategies for retrieving sound clinical studies of the etiology, prognosis, diagnosis, prevention and treatment of disorders in adult general medicine.

In this paper, we describe an automated knowledge extraction method from MEDLINE citations and report on a study of its suitability for providing appropriate concept relationship knowledge. The work combines ideas introduced by Zeng and Cimino with the search strategies by Haynes et al.

METHODS

The UMLS co-occurrence information is stored in the MRCOC table, which is publicly available. Each record contains the UMLS concept unique identifiers (CUI) of the two MeSH concepts that co-occur, the source database (e.g. MEDLINE), the type of occurrence, the number of co-occurrences, and the subheadings that belong to the first concept in each record (and are therefore different for each direction of the relationship). We have built a similar table with citations retrieved by the clinical queries available in PubMed. We have chosen to build our own table because we sought to improve performance by retrieving relationships that were more specific to each clinical task.

The following procedures were used to build the co-occurrence table and to extract the potential relationships from MEDLINE citations:

1. Create a co-occurrence table of MeSH terms from MEDLINE citations using the 4 clinical query categories (therapy, diagnosis, etiology, and prognosis) with emphasis on specificity. For each category, we retrieved the most recent 1000 citations. The subject area was "cardiovascular diseases". The co-occurrence table built was similar to the UMLS MRCOC. We removed the reciprocal entries. [Figure 1]
2. Create a co-occurrence table of semantic types based on the MeSH pairs. For each MeSH pair generated in step 1, we identified the CUIs and collected the corresponding UMLS semantic types. All possible combinations of semantic pairs were created based on each MeSH pair and stored. [Figure 1]
3. Merge entries with the same pair of semantic types into a single entry. During the generation process, information about one pair of semantic types is stored according to the order the semantic pair occurs. It is possible that a relation is stored in both directions. In this study, we considered a pair as having no primary direction. Reversed pairs were merged within a citation (e.g. disease-finding is considered the same pair as finding-disease)
4. Exclude semantic types not relevant to the medical record. Since we were only interested in information that could be potentially relevant to a medical record, pairs containing semantic types that were not relevant to a medical record were excluded. Exclusions were made based on the UMLS documentation description for each semantic type by the researchers. Figure 2 shows a list of a few semantic types included and excluded.
5. A statistical analysis was then performed in order to identify the relevant pairs in each group.

The statistical analysis focused on 2 questions: a) Is the observed relationship between semantic type "X" and semantic type "Y" statistically significant or could the pair occur by chance? b) How strong is the relationship between X and Y?

We performed a chi-square test with Yates correction for each pair generated. Bonferroni correction was used to define the statistical level of significance because of the multiple testing

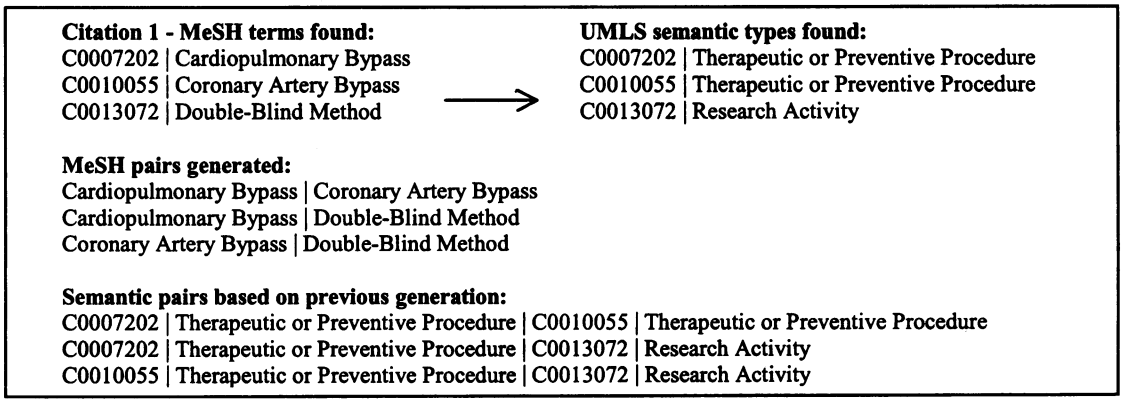


Figure 1. Semantic pairs generation

hypotheses. A phi-coefficient was calculated for each pair.

We also performed a pilot study in order to evaluate the clinical validity of the information retrieved. A questionnaire was designed, which was completed by 5 physicians. Each questionnaire contained 40 pairs of semantic relationships (10 for each clinical category) and examples of MeSH heading pairs that matched to the semantic pair in question. The pairs were randomly selected from the list of pairs generated. A brief explanation of the project was given to the physicians and they were asked whether the selected pairs were relevant to the specific clinical task (see Figure 3).

For each pair, we thus had five different relevance scores based on the physicians' answers. From these scores, we assigned a relevance level to each pair. This relevance level was just the proportion of physicians who indicated the pair as relevant.

To measure the performance of the extraction method, we used physician opinion as a reference standard to assign relevance. However, determining the appropriate relevance level was not straightforward. The relevance requirements for extraction may be sensitive to the clinical task. Extracting only the most relevant information may be optimal for some tasks while including all information that might be relevant is optimal for others. Since it was unknown *a priori* what degree of relevance was optimal for the specific tasks studied, we measured performance at each relevance level.

For each level, sensitivity and specificity of the extraction method were calculated for the different extraction tasks. An estimate of the area under the ROC curve was then computed using the non-parametric A' statistic proposed by Pollack and Norman.[17] These A' values were then averaged across the different relevance levels to calculate a single performance measure for each task. Bootstrapping was used to estimate the variance of this average A' measure.

Types included:
Disease or Syndrome
Congenital Abnormality
Hazardous or Poisonous Substance
Enzyme
Phenomenon or Process
Body Part, Organ, or Organ Component

Types excluded:
Bird
Daily or Recreational Activity
Health Care Related organization
Intellectual Product
Regulation or Law
Professional Societv

Figure 2. Examples of semantic types

If the patient has a **Disease or Syndrome,**
Would you be interested in articles about related **Finding?**
For example:
Multiple System Atrophy | Hypotension, Orthostatic
Valvular heart disease | Body Mass Index
Hypertension, Pulmonary | Scleroderma, Systemic
[ ] Yes [ ] No

Figure 3. Example of a question on therapy

RESULTS

The automated process generated 135,667 MeSH pairs in the therapy group, 110,586 in the prognosis group, 142,915 in the etiology group, and 111,713 in the diagnosis group. The generation of all possible semantic pairs based on the MeSH pairs increased the number of pairs generated. Table 1 shows the number of pairs identified.

The statistical analysis was done after merging the co-occurrence pairs of the same semantic types. Table 1 also shows the number of relationship pairs per group, and the number of unique semantic types that occurred in each group. Note that not all permutations were generated. We found that 157 (6.10%) pairs differ significantly from the others in the therapy group, 161 (7.43%) in the prognosis group, 201 (7.32%) in the etiology group, and 189 (8.51%) in the diagnosis group. ($p < .05$, Bonferroni correction).

The analysis of performance showed that the performance varies across the tasks. Performance in the therapy task was significantly better than in the 3 other tasks ($p<0.05$). Table 1 shows the ROC area for the different extraction tasks. Figure 4 shows the sensitivity and specificity of each task, averaged over the different relevance levels.

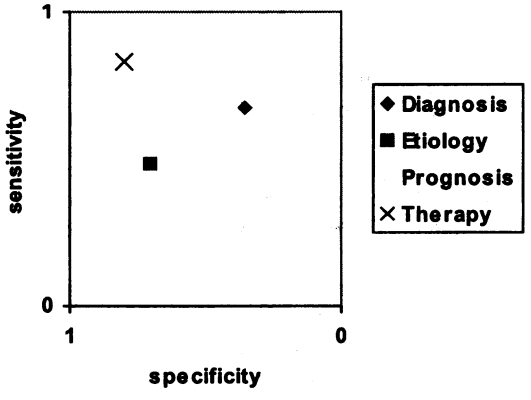
Table 1. Results

Clinical task	MeSH pairs generated	Semantic pairs generated	Semantic pairs after merge	Unique semantic types	Relevant semantic pairs
Therapy	135,667	195,096	2,575	111	157
Prognosis	110,586	141,043	2,168	109	161
Etiology	142,915	188,908	2,745	107	201
Diagnosis	111,713	144,971	2,222	107	189

Table 2. ROC area for the different groups

	Diagnosis	Etiology	Prognosis	Therapy
A'	0.640873	0.650099	0.399107	0.909127
low CI	0.476496	0.500079	0.226024	0.858291
high CI	0.80525	0.80012	0.572191	0.959963

Figure 4. Average Sensitivity & Specificity



DISCUSSION

The primary focus of this experiment was to explore an automated knowledge extraction method to determine its suitability for providing appropriate concept relationship knowledge. The amount of information acquired from a method such this is large. Compared to the amount of time and work required to construct such a knowledge base manually, this process is considerably faster and easier.

The pilot study performed in order to evaluate the clinical validity of the information retrieved showed that the results were suitable for the intended purpose (literature retrieval), especially in the therapy group.

As mentioned in previous studies by Powers[18] et al. and by Cimino and Barnett[8], literature retrieval is one of the potential areas where knowledge extraction can be applied. The semantic relationships identified might serve to improve literature review, producing patterns or rules which may be useful for improving search strategies. For example, a rule can propose that a particular disease or syndrome is related to a certain drug or laboratory test.

Consider the following situation. An elderly patient comes to the hospital complaining of progressively worsening shortness of breath on minimal exertion. The physical exam suggests heart failure. The patient has a history of uncomplicated inferior wall myocardial infarction a few months ago and is taking propranolol. Suppose in this case the clinician needs additional information on the etiology and treatment of the heart failure. The physician may search for "heart failure and therapy", or may use more specific evidence based search strategies such as those by Haynes et. al. Using knowledge such as the relationship between diseases and drugs (see above), the search strategy could be expanded by including propranolol. Propranolol may help to reduce the risk of cardiovascular death in post-MI patients with poor left ventricular function.[19]

We believe that the knowledge generated by the method described in this paper will be particularly useful for the task of retrieving relevant information from the electronic medical record in order to guide the users during the retrieval process and, consequently, improving search strategies and information retrieval.

There are, however, a few concerns regarding the use of information extracted from MEDLINE citations. The relationships generated by this approach are propositions only. In a perfect situation, a medical expert should review the validity of each relationship. Automated extraction of relationships from MEDLINE can produce large quantities of information making a manual review a time-consuming task. The information also depends on the quality of the indexing.

CONCLUSION

The work described in this paper demonstrates that it is possible to extract useful medical knowledge from MEDLINE citations. The amount of information acquired was large although only 7 to 8% of the semantic pairs generated in each task group differ significantly from the others. The pilot study shows a relatively good specificity and sensitivity for the intended purposes of this project. Performance was especially good in the therapy group. The knowledge may not be totally accurate due to the types of errors described. However, we believe that it can serve for the task of retrieving relevant information from the electronic medical record in order to guide the users during the retrieval process

Acknowledgments

The authors thank Adam Wilcox for his assistance in the performance analysis. This work was supported by a Center for Advanced Technology grant from New York State, a Digital Library Initiative grant from the National Science Foundation, and a CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) grant from Brazil.

References

1. Covell DG, Uman GC, Manning PR. Information needs in office practice: are they being met? Annals of Internal Medicine 1985; 103(4):596-9.

2. Timpka T, Ekstrom M, Bjurulf P. Information needs and information seeking behavior in primary health care. Scandinavian Journal of Primary Health Care 1989; 7(2):105-9.

3. Shelstad KR, Clevenger FW. Information retrieval patterns and needs among practicing general surgeons: a statewide experience. Bulletin of the Medical Library Association 1996; 84(4):490-7.

4. Gorman PN, Helfand M. Information seeking in primary care: how physicians choose which clinical questions to pursue and which to leave unanswered. Medical Decision Making 1995; 15(2):113-9.

5. Allen VG, Arocha JF, Patel V. Evaluating evidence against diagnostic hypotheses in clinical decision making by students, residents and physicians. International Journal of Medical Informatics 1998; 51(2-3):91-105.

6. Silberg W.M., Lundberg G.D., Musacchio R.A. Assessing, controlling, and assuring the quality of medical information on the Internet: Caveant lector et viewor--Let the reader and viewer beware. Journal of the American Medical Association 1997; 277(15):1244-5.

7. Barnett GO, Cimino JJ, Hupp JA, Hoffer EP. DXplain. An evolving diagnostic decision-support system. Journal of the American Medical Association 1987; 258(1):67-74.

8. Cimino JJ, Barnett GO. Automatic knowledge acquisition from MEDLINE. Methods of Information in Medicine 1993; 32(2):120-30.

9. Zeng Q, Cimino JJ. Automated knowledge extraction from the UMLS. Chute CG. Proceedings/AMIA Annual Fall Symposium. Philadelphia: Hanley & Belfus Inc., 1998: 568-72.

10. UMLS Knowledge Sources. 10th edition. Bethesda, Maryland: National Library of Medicine, 1999.

11. Pao ML. Concepts of Information Retrieval. Englewood, CO: Libraries Unlimited, 1989.

12. Sackett DL, Richardson WS, Rosenberg W, Haynes RB. Evidence-based Medicine: How to Practice and Teach EBM. New York: Churchill Livingstone, 1997.

13. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. British Medical Journal 1996; 312(7023):71-2.

14. Friedland DJ, Go AS, Davoren JB *et al.* Evidence-Based Medicine. A Framework for Clinical Practice. Stamford, Connecticut: Appleton & Lange, 1998.

15. Hunt DL, Haynes RB, Browman GP. Searching the medical literature for the best evidence to solve clinical questions. Annals of Oncology 1998; 9(4):377-83.

16. Haynes RB, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. Journal of the American Medical Association 1994; 1(6):447-58.

17. Pollack I, Norman DA. A non parametric analysis of recognition experiments. Psychonomic Science 1964; 1:125-6.

18. Powsner SM, Riely CA, Barwick KW, Morrow JS, Miller PL. Automated bibliographic retrieval based on current topics in hepatology: hepatopix. Computers and Biomedical Research 1989; 22(6):552-64.

19. A randomized trial of propranolol in patients with acute myocardial infarction. I. Mortality results. JAMA 1982; 247(12):1707-14.