Generic queries for meeting clinical information needs*†

By James J. Cimino, M.D.
Assistant Professor
Center for Medical Informatics, Department of Medicine

Anthony Aguirre, M.S., M.L.S.
IAIMS Project Head
Augustus C. Long Health Sciences Library

Stephen B. Johnson, Ph.D.
Assistant Professor
Center for Medical Informatics, Department of Medicine

Ping Peng, Ph.D.
Associate Research Scientist
Center for Medical Informatics, Department of Medicine

Columbia University College of Physicians and Surgeons
New York, New York 10032

Center for Medical Informatics
Atchley Pavilion, Room 1310
Columbia-Presbyterian Medical Center
New York, New York 10032

This paper describes a model for automated information retrieval in which questions posed by clinical users are analyzed to establish common syntactic and semantic patterns. The patterns are used to develop a set of general-purpose questions called generic queries. These generic queries are used in responding to specific clinical information needs. Users select generic queries in one of two ways. The user may type in questions, which are then analyzed, using natural language processing techniques, to identify the most relevant generic query; or the user may indicate patient data of interest and then pick one of several potentially relevant questions. Once the query and medical concepts have been determined, an information source is selected automatically, a retrieval strategy is composed and executed, and the results are sorted and filtered for presentation to the user. This work makes extensive use of the National Library of Medicine's Unified Medical Language System℠ (UMLS®): medical concepts are derived from the Metathesaurus, medical queries are based on semantic relations drawn from the UMLS Semantic Network, and automated source selection makes use of the Information Sources Map. The paper describes research currently under way to implement this model and reports on experience and results to date.

INTRODUCTION

Computer-based medical information resources usually respond to users' needs in one of two ways. One approach is to provide a user interface with a variety

* A portion of this paper was presented at the Ninety-second Annual Meeting of the Medical Library Association, May 15–21, 1992, Washington, D.C.
† This research was supported in part by NIH Contract N01-LM-1-3536 from the National Library of Medicine.

of options, each of which offers an answer to a specific question. A common example is a drug information system, in which one command retrieves dosage information, another command retrieves drug interaction information, and so on. Such systems provide high-quality answers to a limited set of questions. Success with such systems depends on the user's information need matching one of the precise questions offered (assuming the user can locate the appropriate system in the first place). At Columbia-Presbyterian Medical Center (CPMC), it has been determined that users consider application "switching" and lack of familiarity with various systems to be an impediment [1].

A second approach is to provide access to a large database and a query language with which to search the database. One well-known example of this method is found in MEDLINE® from the National Library of Medicine (NLM) [2]. Rather than enter commands that correspond to specific questions, users enter search terms that match the literature citations that are expected to provide the desired answer. While such systems are capable of answering a wide variety of questions, success depends on the user formulating the retrieval strategy properly.

Additional strategies are under development. In some experimental systems, a "super resource" provides access to a variety of information sources [3–4]. The super resource offers answers to a large number of specific questions, based on the combined set of possible answers offered by the individual sources. This hybrid approach can offer users the advantages of both of the two more common approaches. The potential exists, however, for overwhelming the user with the sheer number of questions and resources available.

This paper describes two alternative approaches based on a single underlying information model. The model is based on the theory that a majority of user information needs can be mapped to a finite number of general questions, referred to as "generic queries." The authors believe the number of generic queries is small enough to be managed by an information retrieval system but too large to be managed by humans (i.e., a menu of all possible generic queries would overwhelm a user). The two approaches to information retrieval are intended to help the user identify the relevant generic query. One approach makes use of natural language processing, and the other uses what are referred to as "constrained user queries."

The issues addressed in this research include establishing a set of generic queries, processing user queries to match generic queries, making educated guesses about relevant generic queries (based on the context of the user's question), selecting information sources appropriate for the generic queries, mapping user questions to retrieval statements recognized by the appropriate information source, executing the search automatically, and analyzing and presenting retrieval results.

BACKGROUND: THE UMLS

Since its inception in 1986, NLM's Unified Medical Language System™ (UMLS®) project has produced resources to help direct users to appropriate computer-based medical information sources and to assist them with the retrieval process. The principal results of research to date have been a Metathesaurus [5], a Semantic Network [6], and an Information Sources Map (ISM) [7]. The availability of these UMLS resources raises important research questions. For example, how can the UMLS Knowledge Sources improve the mechanism by which a user's information need leads to access to that information? What additional features does the ISM require for automatic selection of and access to information sources? How useful are the UMLS resources and relevant algorithms for improving information access in a variety of end-user settings?

Recent research has helped define the unmet information needs of health care professionals. One study of clinician information needs in doctors' offices suggests that poor organization, inadequate indexing, and the age of information sources, as well as ignorance of their availability, contribute to the failure to satisfy information needs [8]. A survey of general practitioners in Sweden showed frequent, unmet needs for information concerning diagnosis and choice of therapy and suggested that computer technology might support differential diagnosis in general medicine [9]. A study of residents' work rounds, attending physicians' rounds, morning report, and the interns' clinic in a university-based general medicine service revealed the frequent occurrence of patient care questions that potentially could be answered by online information sources [10]. Given all these needs and the potential computer-based solutions, there appears to be a place for mechanisms to facilitate access to clinical information sources. The UMLS Knowledge Sources could play a possible role in these mechanisms; this inspired the current research.

METHODOLOGY

Overview

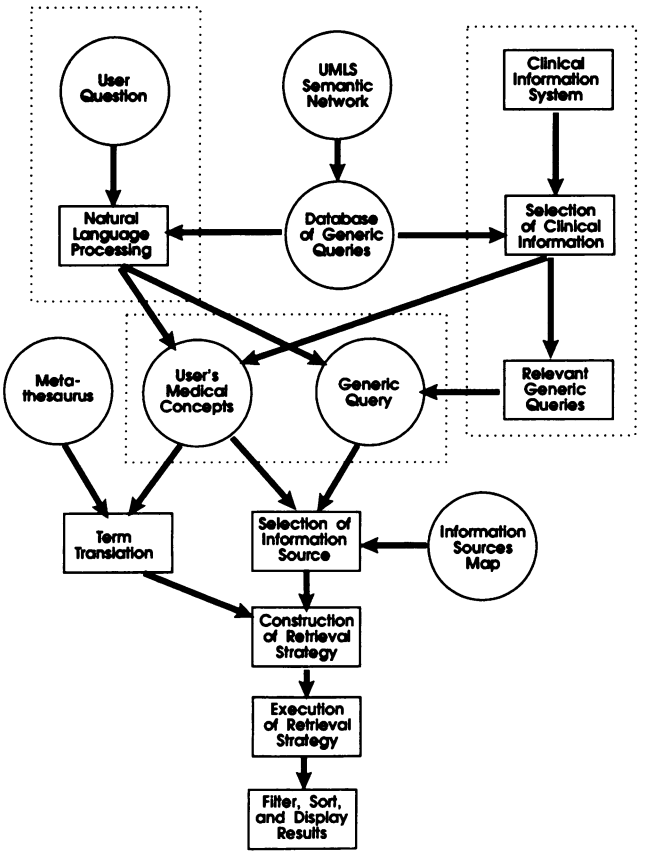
The model is based on the assumption that most user queries for medical information can be matched to one of a finite set of general patterns, or generic queries. By treating individual user queries as combinations of generic questions and sets of specific medical concepts, the authors expect to facilitate the development of more effective systems of locating

and retrieving computer-based information relevant to the original questions.

Figure 1 depicts the various steps involved in the model. Information sources are categorized by the types of generic questions for which they are likely to have answers. Generic queries are converted into formats or templates recognized by the relevant information sources. The specific medical concepts inherent in a particular user question are converted to the vocabulary used by a relevant information source and added to the relevant query template. The enhanced query then can be posed to the information source and the results recorded. These results are processed to detect the information most relevant to the original query. The retrieved results can then be ranked and pertinent parts of each result abstracted for display to the user.

Take, for example, the user question "How effective is DDS for the treatment of Hansen's disease?" This statement is parsed and, through syntactic and semantic analysis, found to match the generic query type "therapy effectiveness." MEDLINE is identified automatically as a likely place to obtain the answer. "Hansen's disease" then is translated to the Medical Subject Heading (MeSH®) term "LEPROSY," and "DDS" is translated to the MeSH term "DAPSONE" (note that semantic information from the parsing of the question ensures that "DDS" will not be mistranslated into "dentist"). From the query template and the MeSH terms, a MEDLINE search strategy is composed to retrieve citations that might address the user's question. Retrieved citations are analyzed and ranked using relevance and quality metrics. For example, citations that mentioned DDS or dapsone frequently and in close proximity to leprosy or Hansen's disease would be ranked high, as would citations that included terms such as *success rate* and *efficacy*. Once the citations are ranked, appropriate abstracted portions are shown to the user, in descending order of relevance. The user has the option of seeing an entire citation or retrieving the full text of an article if the citation seems promising.

Figure 1
Information retrieval using generic queries



Developing generic queries

The crucial first step in this project is to determine a reasonable set of generic queries. While this set need not account for all possible information needs, it must cover a substantial proportion (probably a majority) of the reasons that bring users to computer-based medical information resources.

User information requests for the initial database are being drawn from three sources: a collection of NLM user questions [11], a set of questions related to cystic fibrosis [12], and questions encountered at the reference desk of Columbia University's Health Sciences Library. These questions are analyzed in two ways. Syntactic analysis is carried out through natural language processing techniques to identify the interrelated medical concepts in the query. Semantic analysis is carried out by experienced reference librarians using knowledge acquisition techniques to determine both the nature of the question being asked and the semantic relationships among the medical concepts [13]. As medical concepts are found, they are located in the Metathesaurus, which provides information about the semantic types of the concepts. The UMLS Semantic Network then is examined to determine which (if any) of the semantic relationships in the network reflect those in the query. The result is a reduced form of the question in which the specific concepts are replaced by semantic types, and the meaning of the query is represented by one or more semantic relationships among the semantic types.

In the Hansen's disease example given earlier, the concepts "DDS" and "Hansen's disease" are identified in the Metathesaurus as having the semantic types

"Pharmacologic Substance" and "Disease or Syndrome" respectively. The Semantic Network shows that these two semantic types are associated through the "treats/treated by" relation. The generic query thus would contain the template "⟨Pharmacologic Substance⟩ ⟨treats⟩ ⟨Disease or Syndrome⟩." Of course, many other user questions might map to the same generic query, such as, "Is aspirin the best treatment for headache?" "Is calcium still used for cardiac arrest?" and "Does quinine really work for nocturnal leg cramps?"

Matching user queries to generic queries

A direct approach for determining which generic query matches the user's question is to require the user to select a generic query from the list of all those known to the system. But there are several disadvantages to this approach: the list may be so long that the user is overwhelmed, the generic queries may be so abstract that the user cannot recognize the appropriate one, and the selection of a generic query reveals nothing about the specific medical concepts in the user's question. Therefore, two very different approaches—natural language processing and query constraint—are being examined to achieve the same end; namely, the identification of an appropriate generic query and the specific concepts of interest to the user.

Natural language processing

Question-answering systems are fairly well understood, compared with most other natural language processing systems, and have a substantial literature [14–25]. Such systems are successful enough to be sold commercially, such as INTELLECT 400 of AICORP and Q&A of Symantec Corp [26]. This success is due primarily to the fact that the user is permitted only to pose questions and cannot offer information to the system. By limiting the scope of processing to single queries rather than full text, the natural language processing system is not required to handle as wide a set of syntactic constructs or to deal with the substantial problems of determining relations among sentences (i.e., discourse analysis) [27–28]. When the information resources being queried are databases, document retrieval systems, or simple knowledge bases (fact bases), query understanding is simplified because the domain is formalized, and the system does not require as much general knowledge of the world to understand the user's information needs [29].

Existing question-answering systems cannot process information requests in the medical domain by themselves, because they require knowledge about the particular structure of such requests (a grammar), the terms appearing in the requests (a lexicon), and the relationships and dependencies among the terms (the semantics) [30]. The UMLS offers significant resources for developing a natural language processing system for medical queries. The Metathesaurus and Semantic Network contain a great deal of the knowledge needed for natural language processing, some present explicitly and some derivable through various processing techniques. These resources are expected to evolve over time and to incorporate feedback from many different institutions.

Using the UMLS, the processing system's task of understanding the natural language questions posed by a user can be constrained because the domain of discourse is restricted to biomedicine. This greatly reduces the sheer size of the processing system and minimizes problems of ambiguity because knowledge about the semantics of queries can be built into the system [31–35]. The authors' approach does not require complete semantic analysis of the question. All that is required is sufficient syntactic and semantic analysis to match the question to one of the predetermined generic queries and to identify the relevant medical concepts. Using the initial database of user questions, a linguistic database is being compiled that includes word frequencies, word and phrase cooccurrence statistics, an inventory of syntactic patterns for the queries, determination of semantic categories of words and phrases, and identification of semantic patterns [36–37].

The natural language processing system for interpreting user questions makes use of three principal databases of linguistic information: a Lexicon, a Grammar, and Semantic Patterns. The Lexicon provides syntactic and semantic information for each word and idiom (i.e., phrase that cannot be analyzed into smaller units) in the user queries and the UMLS Knowledge Sources. The Grammar is a set of rules that specify the syntactic structure of phrases and full questions, using the syntactic categories and other properties defined in the Lexicon. For example, *pleural effusion* might be assigned a syntactic structure such as NP ([N "effusion"] [ADJ "pleural"]), indicating a noun phrase (NP) consisting of the noun *effusion* with modifying adjective *pleural*. Semantic Patterns defines how single semantic categories combine into larger semantic units. For example, *pleural effusion* might be assigned the pattern "BODY-PART+ MORPHOLOGY." Semantic Patterns reduces ambiguity by specifying combinations of semantic categories that are permitted [38–39].

The interpretation algorithm consists of two main steps: the processing of the user query into a semantic structure and the mapping of this semantic structure to the appropriate query template. Two methods are being explored for the first step: syntactic parsing and semantic pattern matching. In syntactic parsing, the syntactic structure of the user query is obtained and translated into a semantic structure [40]. Semantic pat-

tern matching identifies semantic patterns directly from a user query without using syntactic information. Syntactic parsing is expected to perform effectively with complex queries but less well with input that is poorly formed syntactically; conversely, semantic pattern matching is expected to work better when the user query is poorly formed than when queries are long and complex.

After a semantic structure is obtained, a template-mapping algorithm is used to produce a query template. The algorithm first determines which stored template has the closest fit with the query and then fills the "slots" of the template with appropriate information extracted from the query [41].

Consider the following user query taken from the library database: "Identify papers referring to laryngeal cancer giving metastasis to mandible." A query interpretation system must be able to handle queries phrased as commands rather than as questions. The system would determine that the user wants a literature search because of the request for papers. The system may need to know that *neoplasm* and *tumor* can be synonyms for "cancer," and that *jaw* can be a synonym for "mandible." The system should understand the semantic structure of the query; that is, "disease cancer metastasizes from location larynx to location mandible," where the relation "metastasizes" relates the disease term "cancer" to the anatomic terms "larynx" and "mandible." If information retrieval systems can be designed to make use of this contextual information, then the precision and recall of queries probably can be improved [42]. For example, given the semantic structure for this query and a knowledge base of anatomy, the interpretation system might be able to generate disease terms more specific than "laryngeal cancer," such as *glottic tumor, supraglottic tumor,* and *subglottic tumor.*

Matching by query constraint

In some situations, a system should be able to determine a user's information need without resorting to natural language processing. Consider a user of a clinical information system who is reviewing patient data. Suddenly, during review of a particular set of information in a particular clinical application or context, a question arises. Certain inferences can be drawn. First, the information needed involves in some way the specific patient data under review, and those data likely would be involved in a query to satisfy the need. Second, for any given context in which a need arises, there is probably some relatively small set of typical needs corresponding to generic queries. For example, a user reviewing laboratory data likely would have questions about the differential diagnosis of a laboratory abnormality or the predictive value of a laboratory procedure.

The CPMC Integrated Advanced Information Management System (IAIMS) project includes the Presbyterian Hospital Clinical Information System (CIS), which provides access to a variety of coded clinical data, including laboratory results, in-patient diagnoses and procedures, and outpatient problems and medications [43]. The in-patient terms are encoded using the *International Classification of Diseases, 9th Revision, Clinical Modifications, (ICD-9CM)* [44]; the laboratory data are encoded under a local vocabulary; and the out-patient terms are encoded using a mixture of the *Systematized Nomenclature of Medicine (SNOMED)* [45] and local terms.

In any given application or context, the patient information displayed is of a very small number of semantic types (usually one or two), similar to those in the UMLS Semantic Network and in the generic queries. By identifying the queries with semantic types matching those in a particular context, the set of relevant queries can be limited in size. When the number of queries is greater than one, the user can select the most relevant query. Because the medical concepts are drawn from the information system context, the queries can be transformed from generic forms to specific forms more familiar to the user. For example, when reviewing laboratory test results, the queries would address questions about differential diagnosis of abnormal results, test characteristics (i.e., sensitivity, specificity, positive predictive value), causes of spurious results, complications, and contraindications. When reviewing a patient problem, the queries might address questions about diagnostic tests, differential diagnosis, therapies, and contraindications.

Information source selection

The automatic selection of an information source based solely on the results of natural language processing would require a machine with capabilities that do not yet exist. However, selection of an information source for a generic query is a much more tractable problem.

In some cases, the selection can be predetermined, based on the particular query and the available sources. For example, a query of the form "What is the correct dose for drug X?" probably could be served best by a drug database, such as an electronic version of the *Physician's Desk Reference.* In other cases, some procedural component may be needed, based on the concepts identified in the query. For example, for a query of the form "What is the latest treatment for disease X?" MEDLINE might be the most appropriate database. However, depending on what X is, more appropriate sources may be available (e.g., if X is a cancer, NLM's PDQ®). The selection process can be straightforward if each generic query can be assigned one or more appropriate sources in advance. The more

difficult question is how those assignments should be made.

In this project, reference librarians select information sources for sample user questions and then attempt to determine the important characteristics of the questions that influenced the choices. Based on this information, the syntactic and semantic aspects of the questions (and the corresponding generic queries) that appear to influence the selection of a particular source are determined. These associations can then be generalized into a pattern-matching process that can be applied to other user questions and generic queries.

An important aspect of the development of the pattern-matching process is to seek attributes in the ISM that can assist in the proper source selection. For example, if the query includes the semantic relationship "drug treats disease," the ISM records can be searched to identify sources that include information about drugs, diseases, and therapy. As the ISM is still in an early stage of development, the present research may determine what *should be*, rather than what *is* in the ISM for the purpose of automated source selection.

Mapping user questions to retrieval statements

Once the generic query, specific concepts, and information source have been identified, a query must be constructed to yield appropriate results when applied to the information source. The two general parts to this problem are translation (if necessary) of the user's concepts into those recognized by the information source and construction of a query statement or a command sequence.

Processing identifies terms in the lexicon that correspond to medical concepts in the query. Each concept in the lexicon is mapped in advance to one or more Metathesaurus concepts. For most information sources, these concepts can be left in their original form. When the source (e.g., MEDLINE) uses some particular vocabulary contained in the Metathesaurus (e.g., MeSH terms) and the original terms are not in that vocabulary, retrieval may be improved by mapping the original concepts to terms in the source vocabulary. This mapping is carried out by identifying source terms that are lexical variants, synonyms, or otherwise related in the Metathesaurus to the original terms. Suppose, for example, that the user's concept is "Hansen's disease" and the desired information source is keyed to MeSH terms. The Metathesaurus includes "HANSEN'S DISEASE," which is linked as a synonym to the term "LEPROSY." The Metathesaurus also shows that "LEPROSY" is a MeSH "preferred" term (i.e., used for indexing citations).

Each generic query is associated with an appropriate command sequence or expression for use in executing a retrieval strategy. The strategy is designed for use with a particular information source and contains one or more place holders for the user's concepts. For example, the question "How do I treat Hansen's disease?" might become the generic query "What is treatment for ⟨disease⟩?" If MEDLINE were the information source associated with the generic query, the command sequence might be "⟨X⟩ with Drug Therapy." When retrieving information, the system translates the user's concept into MeSH terms and inserts the result into the command sequence to produce the search strategy "LEPROSY WITH DRUG THERAPY."

Execution of the retrieval strategy

Execution of the retrieval strategy requires a method for accessing the desired source, transmitting the strategy, and capturing the results. One technical solution applicable to some sources is to use a personal computer with terminal emulator software and a scripting facility to establish a remote look-up to a computer running the source program. Other approaches include the use of application program interfaces, such as those used by the Wide-Area Information Servers, to send information requests over a computer network [46].

Analysis and display of results

In most cases, information retrieval involves a tradeoff between thoroughness and precision: increasing one usually sacrifices the other. One way to balance these opposing qualities is to perform a broad search (i.e., one sensitive to appropriate information but that also may pick up a large amount of irrelevant material) and then to rank the results by relevance. Relevance would be determined through some postprocessing technique not available in the original source. Some available retrieval systems are capable of carrying out this process. This approach can be generalized to other sources. For example, a search on a MEDLINE database can retrieve all citations with a particular keyword but usually cannot retrieve all citations in which the keyword is repeated. This result can be achieved by performing a search for the keyword and then excluding citations that contain the word only once.

A simple count of a keyword is but one heuristic for measuring the quality of retrieved results. Published guidelines exist for determining the quality of the medical literature. Sackett et al. describe basic criteria that might be required to establish quality, such as a specific description of the setting in which the study or treatment was carried out, selection methodology, numbers, and sociodemographic factors of the participants [47]. Criteria also are suggested

for specific types of studies, such as those dealing with therapeutics, diagnosis, prognosis, and etiology. Riegelman and Hirsch describe methods for evaluating test procedures for developing quality tokens [48]. Measures such as these, as well as the use of positional operators [49] and other weighting techniques [50–52] can be applied to medical text retrieval. For example, relevance feedback in combination with term weighting has been shown to be useful for filtering retrieved material [53] and predicting system performance [54]. Haynes suggests words or phrases corresponding to these concepts that could be sought in the retrieved text so that material could be ranked by the frequency of their occurrence [55]. For the concept of diagnosis, for example, such terms include *criterion standard, gold standard,* and *blinded or masked comparison.*

CURRENT STATUS OF PROJECT

At this writing, the first year of the three-year project has been completed. The initial effort has focused primarily on the development of the methodology: several portions of the research plan have been implemented.

Fifty-three user questions from the NLM set were analyzed by two experienced reference librarians. They identified main concepts and expressed relations between concepts in their own terms. Queries (such as, "How is furosimide used for high blood pressure?") were broken down into triples of the form concept-relation-concept (such as, "furosimide—used for—high blood pressure"), resulting in sixty-eight triples. After mapping the concepts to the Metathesaurus and the relations to the UMLS Semantic Network, the sixty-eight triples could be expressed with nineteen relations and forty semantic types (e.g., "pharmacologic substance—treats—diseases or syndrome" for the above example). Linguistic analysis produced more detailed results, obtaining more triples per query. Many of these triples reflect background information related to the user's query rather than the primary focus of the query ("core" triples). Although there is some linguistic basis for determining the query focus, there is not yet sufficient evidence that the focus can be determined automatically by natural language processing programs.

The Lexicon, Grammar, and Semantic Rules components have been developed to cover most of the essential structures found in the user queries in the NLM test collection. However, these components are far from complete. Much of the Semantic Rules can be extracted from the UMLS Metathesaurus and Semantic Network, but manual review and entry of additional information is time consuming. Some significant problems remain in interpreting user queries: handling complex forms of conjoined phrases (i.e., phrases connected by *and, or,* or a comma); resolution of pronouns; establishing connections between sentences; and reconstructing material omitted by the user (usually for brevity).

Generic queries developed thus far have been integrated with a clinical application [56]. The clinical application (the Admission Profile) displays *ICD-9CM* diagnoses and procedures associated with patient admissions. Users may select one or two *ICD-9CM* terms (Figure 2). The generic queries were examined to determine those of potential relevance to a user. Because the Admission Profile presents the user with disease and procedure terms, generic queries were selected that included "disease or syndrome," "pathologic function," "injury or poisoning," "mental or behavioral dysfunction," "congenital abnormality," "acquired abnormality," "health care activity," "diagnostic procedure," "laboratory procedure," "therapeutic or preventive procedure," or "biomedical occupation or discipline." Forty potentially relevant queries were found; English-language questions based upon these queries were constructed. In this process, several of the queries were merged into single questions, resulting in a total of 18 questions, 8 involving a single disease, 2 involving two diseases, 4 involving a single procedure, 1 involving two procedures, and 3 involving one disease and one procedure.

When a user selects one or two *ICD-9CM* terms, the system selects questions (contained in CPMC's *Medical Entities Dictionary*) that are particularly relevant to the types and numbers of selected terms. For example, if the user selects one disease term and one procedure term, generic queries are presented that have place holders for these types of terms. The system generates English-language versions of the generic queries by translating the *ICD-9CM* terms to MeSH terms, inserting them into the query, and displaying them to the user (Figure 3). When the user selects one of these questions, a search strategy is generated (Figure 4). In this application and for these queries, the selected source is a five-year set of MEDLINE citations using the BRS/Onsite search engine. The system passes the search strategy to the search engine and the retrieved results are displayed using the standard BRS interface.

The system is operational but cannot be tested until adequate translation of *ICD-9CM* to MeSH is possible. At present, 48% of *ICD-9CM* terms can be converted to MeSH terms, although in most cases the translations are crude at best [57].

DISCUSSION

Other researchers have developed systems designed to anticipate users' information needs so as to direct users to appropriate information sources and to fa-

Figure 2
Admission Profile screen with two terms selected (X)

```
 Name: SMITH, L              Sex: F  Birthdate: 10/07/909   MRN: 0963210
                          Admission Record Detail
-------------------------------------------------------------------------
   Admission Date:  06/06/92    Discharge Date:  06/12/92   Location:  M6HN
    Doctor:  FELDTWEIS, MICHAEL          Discharge Summary:  Y

   Select Terms You Are interested in:

         Diseases:
        097.1    LATENT SYPHILIS NOS
        170.4    MAL NEO LONG BONES ARM
    X   423.0    HEMOPERICARDIUM
    X   427.2    PAROX TACHYCARDIA NOS
        345.90   EPILEPSY, UNSPECIFIED,W/O INTR

         Procedures:
        03.31    SPINAL TAP
        37.5     HEART TRANSPLANTATION


-------------------------------------------------------------------------
 Help=F1  Search MEDLINE=Enter          Discharge Sum. List=F5
Admission List=F3 Scroll Down=F8 Prev Test=F9 CIS Main Menu=F12 Signoff=F11
```

cilitate retrievals [58–59]. The present model uses natural language analysis to determine the underlying information needs. Previous work has classified information needs to a degree where manual breakdown of the question can facilitate information retrieval [60]. The present methods are intended to analyze user questions to a degree that permits automated classification, automated selection of information sources, and automated retrieval.

The "divide and conquer" method of separating the concepts from the query type is an important aspect of this approach. One advantage of the separation is that it reduces the number of queries that might be encountered from potentially infinite to more manageable. Another advantage lies in the semantic information available as a result of the separation process. When terms are translated from one vocabulary to another, ambiguities can make the process difficult. Semantic information deepens understanding of relationships between the terms and their queries and thus can clarify translation ambiguities (such as whether to translate "DDS" to "dapsone" or "dentist," as in the earlier example).

The separation of query from term is a process that parallels the division of knowledge in the UMLS. Given the information sources available today, it is the query type rather than the query content that determines the appropriate information source. For instance, a query about disease manifestations, regardless of the disease, might be served better by a medical diagnosis program than by MEDLINE. Similarly, the ISM deals with the types of information available in various medical resources, without going into all the detailed concept knowledge. The ISM thus is better suited to respond to a generic rather than a specific question. The approach outlined in this paper will help identify the generic questions that users have and how these questions are or can be represented in the ISM.

The separation of terms from queries also parallels the notion of a Metathesaurus. The Metathesaurus is a resource that can be used to translate terms from one vocabulary to another. It is not organized to translate queries, only the terms they contain, and, therefore, it is necessary to isolate the terms from the query prior to translation. For example, an attempt to translate a query such as, "What is the latest stomach cancer therapy?" would be difficult, unless the terms "stom-

Figure 3
Generic queries based on the terms selected in Figure 2

```
Name: SMITH, L              Sex: F  Birthdate: 10/07/909   MRN: 0963210
                MEDLINE Queries from Admission Profile
--------------------------------------------------------------------------

Select a question:


1.   Does Pericardial Effusion involving Hemorrhage cause Tachycardia,
     Paroxysmal?

2.   Does Tachycardia, Paroxysmal cause Pericardial Effusion involving
     Hemorrhage?

3.   Do Tachycardia, Paroxysmal and Pericardial Effusion involving
     Hemorrhage occur together?




--------------------------------------------------------------------------
Help=F1        Search MEDLINE=Enter
               Admission Detail Menu=F3  CIS Main Menu=F12  Signoff=F11
```

ach cancer" and "therapy" were identified. Attempting to translate the query as a whole might result in translations of *stomach* and *cancer therapy*, which would be less likely to address the user's question. The approach presented in this paper will help identify the medical terms that appear in user queries and how they are or can be represented in the Metathesaurus.

It also will be important to keep track of the relationships among terms in the query. For instance, the query, "Does hyperlipidemia cause myocardial infarction?" can be reduced to the terms "hyperlipidemia" and "myocardial infarction" and to the generic query "does ⟨disease⟩ cause ⟨disease⟩?" Awareness of the actual semantic relationships will help avoid posing the question as "Does myocardial infarction cause hyperlipidemia?" The UMLS Semantic Network is designed to contain precisely this sort of information. The approach outlined in this paper will help identify the semantic relationships that appear in user queries and how these relationships are or can be represented in the Semantic Network.

The user interaction required for constrained user queries will be designed with the same look and feel as the present, familiar applications. In addition, users have demonstrated a willingness to perform MEDLINE searches manually; the constrained user query function will provide similar access through a much simpler interaction. Finally, experience with evaluating other systems has shown that a simple entrance question followed by a simple exit question can be quite effective in eliciting significant user feedback [61-62]. The same approach will be used with the constrained query function.

Initial results suggest answers to some of the questions about the validity of our approach. First, is there a manageable set of generic questions that can represent a large proportion of user information needs? Preliminary studies have revealed recurring semantic relationships in questions asked of reference librarians. This is encouraging, but further study is needed to determine whether, as the number of questions under study increases, the number of patterns will continue to grow proportionally or will stabilize at some manageable number.

Second, if a reasonable number of representative queries are identified, can users' questions be matched to the appropriate queries? In regard to natural language processing, success will depend on the capa-

Figure 4
Search strategy based on the query selected in Figure 3

```
 Name: SMITH, L              Sex: F  Birthdate: 10/07/909   MRN: 0963210
                    BRS Query from Admission Profile
--------------------------------------------------------------------------


 Pericardial Effusion WITH CO AND Hemorrhage WITH ET AND Tachycardia,
 Paroxysmal









--------------------------------------------------------------------------
 Help=F1       MEDLINE Queries=Enter/F3
                                        CIS Main Menu=F12  Signoff=F11
```

bility of the lexicon and the adequacy of the syntactic and semantic patterns accrued. Thus far, the Metathesaurus has proven to be a fairly rich lexicon and the patterns have been fairly easy to acquire. In the case of constrained queries, given a set of reasonable queries, the answer is yes, because the user selects the query.

Third, given that a user's information need can be represented by a generic query, can the query be translated properly for use in information retrieval? Preliminary work with *ICD-9CM*-to-MeSH translation has revealed the severe limitations of current capabilities. However, NLM and its contractors continue to work toward the stated UMLS goal of intervocabulary translation.

The fourth and ultimate question is, can information retrieval be enhanced through the use of generic queries? Retrieval and display strategies for a fixed set of questions can be created in advance by experienced librarians. It seems intuitive that these strategies should be superior to those carried out by a typical user, who must not only devise the strategy but must also select an appropriate information source. This project eventually will attempt to verify this hypothesis.

CONCLUSION

This paper describes a unique approach to satisfying clinical information needs. The clinician's needs are matched to one of a set of general query types for which retrieval strategies have been developed in advance. Matching is accomplished either by processing natural language articulations of the needs or by providing lists of queries that anticipate the needs. Initial work on the various system components needed to execute this approach has been encouraging. The UMLS Knowledge Sources are well positioned to facilitate the processes involved in this approach; the authors' work will expand the UMLS to enhance this effect.

ACKNOWLEDGMENTS

The authors thank the *Bulletin* reviewers and especially Betsy Humphreys and George Hripcsak for

editorial assistance in the preparation of this manuscript.

REFERENCES

1. SUMMERS L, COLOMBOTOS J. Summary of field evaluation of IAIMS. February 19, 1991. (Columbia University internal IAIMS report, available upon request.)
2. HAYNES RB, MCKIBBON KA, WALKER CJ, RYAN N ET AL. Online access to MEDLINE in clinical settings. A study of use and usefulness. Ann Intern Med 1990;112(1):78–84.
3. CIMINO C, BARNETT GO. Standardizing access to computer-based medical resources. In: Miller RA, ed. Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care, November 4–7, Washington, DC. New York: IEEE Computer Society Press, 1990:33–7.
4. MILLER PL, PATON JA, CLYMAN JI, POWSNER SM. Prototyping an institutional IAIMS/UMLS information environment for an academic medical center. Bull Med Libr Assoc 1992 Jul;80(3):281–7.
5. TUTTLE M, SHERERTZ D, OLSON N, ERLBAUM M ET AL. Using Meta-1—the 1st version of the UMLS Metathesaurus. In: Miller RA, ed. Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care, November 4–7, 1990, Washington, DC. New York: IEEE Computer Society Press, 1990:131–5.
6. MCCRAY AT, HOLE WT. The scope and structure of the first version of the UMLS Semantic Network. In: Miller RA, ed. Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care, November 4–7, Washington, DC. New York: IEEE Computer Society Press, 1990:126–30.
7. LINDBERG DAB, HUMPHREYS BL. The UMLS knowledge sources: tools for building better user interfaces. In: Miller RA, ed. Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care, November 4–7, Washington, DC. New York: IEEE Computer Society Press, 1990:121–5.
8. COVELL DG, UMAN GC, MANNING PR. Information needs in the office practice: are they being met? Ann Intern Med 1985;103(4):596–9.
9. TIMPKA T, EKSTROM M, BJURULF P. Information needs and information seeking behavior in primary health care. Scand J Prim Health Care 1989;7(2):105–9.
10. OSHEROFF JA, FORSYTHE DE, BUCHANAN BG, BANKOWITZ RA ET AL. Physicians' information needs: analysis of questions posed during clinical teaching. Ann Intern Med 1991; 114(7):576–81.
11. SCHUYLER PK, MCCRAY AT, SCHOOLMAN HM. A test collection for experimentation in bibliographic retrieval. In: Barber B, Cao D, Qin D, Wagner G, eds. MEDINFO 89: proceedings of the Sixth Conference on Medical Informatics, Beijing, China, October 16–20 and Singapore, Republic of Singapore, December 11–15. Amsterdam: North-Holland, 1989:910–12.
12. SHAW WM, WOOD JB, WOOD RE, TIBBO HR. The cystic fibrosis database: content and research opportunities. Libr Inf Sci Res 1991;13:347–66.
13. MCGRAW KL, HARBISON-BRIGGS K. Knowledge acquisition principles and guidelines. Englewood Cliffs, NJ: Prentice Hall, 1989.
14. GRISHMAN R, HIRSCHMAN L. Question answering from natural language medical databases. Artif Intell 1978;11(1-2):25–43.
15. GROSZ BJ, APPELT D, MARTIN P, PEREIRA F. TEAM: an experiment in the design of transportable natural language interfaces. Artif Intell 1987;32(2):173–244.
16. HARRIS LR. ROBOT: a high performance natural language interface for data base query. In: Bolc L, ed. Natural language based computer systems. Munich: Hanser, 1980: 285–318.
17. HAYES PJ, CARBONELL JG. Multi-strategy construction—specific parsing for flexible database query and update. In: Sridhasan NS, ed. Proceedings of the Sixth International Joint Conference on Artificial Intelligence, Detroit, MI, August 20–25. Menlo Park, CA: American Association of Artificial Intelligence, 1981:432–9.
18. HENDRIX G, SACERDOTI E, SAGALOWICZ D, SLOCUM J. Developing a natural language interface to complex data. ACM Trans Database Syst 1978;3(2):105–47.
19. KAPLAN SJ. Cooperative responses from a portable natural language database query system. In: Brady M, Berwick RC, eds. Computational models of discourse. Cambridge, MA: MIT Press, 1983:167–208.
20. KITTREDGE RI. Natural language queries for a linguistic database using PROLOG. In: Goebel R, ed. Proceedings of the third biennial conference of the Canadian Society for Computational Studies of Intelligence, June 6–10, Montreal, Quebec, Canada. San Mateo, CA: Morgan Kaufman, 1980: 151–7.
21. LEHRERT W. A conceptual theory of question answering. In: Proceedings of the Fifth International Joint Conference on Artificial Intelligence, June 6–10. Cambridge, MA: William Kaufman, 1977:158–64.
22. MARTIN P, APPELT D, PEREIRA F. Transportability and generality in a natural language interface system. In: Proceedings of the Eighth International Joint Conference on Artificial Intelligence, September 3–7, Karlsruhe, West Germany. San Mateo, CA: Morgan Kaufman, 1983:573–81.
23. WARREN DH, PEREIRA FC. An efficient easily adaptable system for interpreting natural language queries. Comput Ling 1982;8(3-4):110–22.
24. WEBBER BL. Questions, answers, and responses. In: Mylopoulos J, Brodie M, eds. On knowledge base systems. New York: Springer Verlag, 1986:400–17.
25. WOODS WA. Lunar rocks in natural English: explorations in natural language question answering. In: Zampoli A, ed. Linguistic structures processing. New York: Elsevier North-Holland, 1977:521–69.
26. OBERMEIER KK. NLIs lead the way. DEC Prof 1988 May; 7(5):70–78.
27. ALLEN J. Natural language understanding. Menlo Park, CA: Benjamin/Cummings, 1987:469.
28. JOHNSON SB. Temporal information in medical narrative. In: Sager N, Friedman C, Lyman M, eds. Medical language processing—computer management of narrative data. New York: Addison-Wesley, 1987:175–94.
29. ALLEN, op. cit.
30. BOGURAEV B, BRISCOE T, EDS. Computational lexicography for natural language processing. New York: Wiley, 1989.
31. JOHNSON SB. Mathematical building blocks. AI Expert 1987 May;2(5):42–50.

32. JOHNSON SB. Review of the form of information in science. Comput Ling 1989;15(3):109–21.
33. JOHNSON SB, GOTTFRIED M. Sublanguage analysis as a basis for a controlled medical vocabulary. In: Kingsland LC, ed. Proceedings of the Thirteenth Annual Symposium on Computer Applications in Medical Care, November 5–8, Washington, DC. New York: IEEE Computer Society Press, 1989:519–23.
34. SAGER N, FRIEDMAN C, LYMAN M, CHI E ET AL. The analysis and processing of clinical narrative. In: Salamon R, Blum B, Jorgensen M, eds. MEDINFO 86: proceedings of the Fifth Conference on Medical Informatics, Washington, D.C., October 26–30. Amsterdam: North-Holland, 1986: 1101–5.
35. SAGER N, FRIEDMAN C, LYMAN M. Medical language processing—computer management of narrative data. New York: Addison-Wesley, 1987.
36. JOHNSON. Mathematical building blocks.
37. SAGER. Medical language processing.
38. JOHNSON. Mathematical building blocks.
39. SAGER. Medical language processing.
40. JOHNSON SB. An analyzer for the information content of sentences [Dissertation]. New York University, 1987.
41. SAGER. Medical language processing.
42. SALTON G, MCGILL MJ. Introduction to modern information retrieval. New York: McGraw-Hill, 1983:258.
43. HENDRICKSON G, ANDERSON RE, CLAYTON PD, CIMINO JJ ET AL. The integrated academic information management system at Columbia-Presbyterian Medical Center. MD Comput 1992;9(1):35–42.
44. UNITED STATES NATIONAL CENTER FOR HEALTH STATISTICS. International classification of diseases, 9th revision, clinical modifications, ICD-9CM. 2d ed. U.S. Department of Health and Human Services, Public Health Service, Healthcare Financing Administration. Washington, DC: 1980.
45. COTE RA, ED. Systematized nomenclature of medicine. 2d ed. Skokie, IL: College of American Pathologists, 1982.
46. DYSON E. WAIS: many ways to do it (wide-area information servers). Release 1.01991;91(4):7–8.
47. SACKETT D, HAYNES RB, TUGWELL P. Clinical epidemiology: a basic science for clinical medicine. Boston: Little Brown, 1985.
48. RIEGELMAN RK, HIRSCH RP. Studying a study and testing a test. 2d ed. Boston: Little Brown, 1989.
49. MCKININ EJ, SIEVERT ME, JOHNSON ED. Using repetition to increase precision in files with large blocks of text. Online Rev 1989;13:369–82.
50. FRISSE ME. Searching for information in a hypertext medical handbook. Commun ACM 1988;31(7):880–6.
51. HERSCH WR, GREENES RA. SAPHIRE—an information retrieval system featuring concept matching, automatic indexing, probabilistic retrieval and hierarchical relationships. Comput Biomed Res 1990;23(5):410–25.
52. EVANS DA, HERSH WR, MONARCH IA, LEFFERTS RG ET AL. Automatic indexing of abstracts via natural language processing using a simple thesaurus. Med Decis Making 1991;11(4 suppl.):108–15.
53. ROBERTSON SE, THOMPSON CL, MACASKILL MJ. Weighting, ranking, and relevance feedback: a front-end system. J Inf Sci 1986;12:71–5.
54. LOSEE RM. An analytic measure predicting information retrieval system performance. Inf Proc Manage 1991;27(1): 1–13.
55. HAYNES RB, MULROW CD, HUTH EJ, ALTMAN DG ET AL. More informative abstracts revisited. Ann Intern Med 1990; 113(1):69–76.
56. CIMINO JJ, JOHNSON SB, AGUIRRE A, RODERER N ET AL. The MEDLINE button. In: Frisse ME, ed. Proceedings of the Sixteenth Annual Symposium on Computer Applications in Medical Care, November 8–11, Baltimore, MD. New York: McGraw-Hill, 1992:81–5.
57. CIMINO JJ, BARROWS RC, ALLEN B. Adapting ICD-9CM for clinical decision support [Abstract]. In: Ackerman MJ, Musen MA, Cox J, eds. Proceedings of the Second Annual Educational and Research Conference, May 6–9, Portland, OR. Bethesda, MD: American Medical Informatics Association, 1992:34.
58. CIMINO. Standardizing access.
59. MILLER, op. cit.
60. FLORANCE V. Medical knowledge for clinical problem solving: a structural analysis of clinical questions. Bull Med Libr Assoc 1992 Apr;80(2):140–9.
61. CIMINO JJ, BARNETT GO, HOFFER EP, PACKER MS ET AL. Building the DXplain knowledge base [Abstract]. In: Jorgens J, ed. Proceedings of the Association for the Advancement of Medical Instrumentation, twenty-third annual meeting and exposition, May 14–18, Washington, DC. 1988: 24.
62. RODERER NR, BARNES S, JANKOWSKI TA. Evaluation report: Compact Cambridge MEDLINE. In: Woodsmall RM, Lyon-Hartmann B, Siegel E, eds. MEDLINE on CD-ROM. Medford, NJ: Learned Information, 1989:117–215.

Received August 1992; accepted September 1992