

Empirical Formulation of a Generic Query Set for Clinical Information Retrieval Systems

Russell J. Cucina^a, Maulik K. Shah^a, Daniel C. Berrios^{b,a}, Lawrence M. Fagan^a

^aStanford Medical Informatics, Department of Medicine, Stanford University, Stanford, California, USA

^bVA Palo Alto Health Care System, Palo Alto, California, USA

Abstract

Information needs in clinical practice take the form of specific questions about a given clinical situation, and are best satisfied by concise and specific information retrieval. We sought to develop a comprehensive set of generic queries for information retrieval from electronic medical information resources. We collected one hundred and ten real-world questions asked at the point of care in a variety of settings, and from these developed a set of generic queries of which each of the real-world queries could be shown to be a special case. To provide allowed values for each of the concept terms in the queries, we defined generic nouns as unions of UMLS semantic types, and specified which of these were appropriate to each query. We have begun to use the set to index reference texts from general and subspecialty medicine, and found it capable of full text indexing in the clinical domain. We hypothesize that the query set can serve as a basis for more specialized query sets, and that it will remain generalizable to other electronic medical resources, indexing tasks, and non-UMLS controlled vocabularies.

Keywords:

Information Storage and Retrieval; Information Systems; Medical Informatics Applications; Decision Making, Computer-Assisted; Medical Informatics Computing; Medical Informatics

Introduction

Electronic clinical information resources continue to grow in availability, and are an increasingly important reference for practicing clinicians. On-line resources previously consisted primarily of citation information or abstracts, such as Medline. Full-text information is now widely available on the Internet, including medical reference texts, the full text of major journals, and online content designed primarily for the Web. With the increased complexity of available information, clinicians require new methods of full-text information retrieval.

A key determinant of an information retrieval system's performance is the user's ability to convert their information need into a query understood by the system. Systems using Boolean linkage of search terms have been shown to be poorly utilized by many users [1], although several semi-automated or automated systems have been developed [2]. Natural language query systems are intuitive to use, but translating meaning into a structured query is difficult to accomplish with accuracy and completeness. An intermediate approach is to define a set of generic queries that specify the relationships among search terms, with the user providing values for the terms to complete the query.

Several authors have explored combining search terms with semantic relationships in information retrieval [3-8]. Miller and colleagues [3] enumerated 34 semantic relationships between generic concepts within the domain of liver disease, and used these relationships to classify a test set of Medline abstracts. They demonstrated the potential for semantic relationships to partition the clinical literature. Cimino and colleagues [4] reported the analysis of 40 complex user questions posed to reference librarians, which they used to develop 37 generic queries with a "noun – relation – noun" structure. They used the relational information encoded in the query structure to construct generic Medline search strategies, using MeSH subheadings to express the relationships between search terms. Ely and colleagues [8] performed an observational analysis of the information needs of 103 family physicians in an outpatient setting. They collected 1101 specific questions, which they grouped into 69 generic categories. They reported the 10 most common in the form of generic queries containing one or two concepts and a semantic relationship.

In contrast to prior work, we sought to develop a query model suitable for both full-text indexing and retrieval, requiring a more compact query set to keep the indexing task manageable. We empirically developed a set of generic queries [Figure 1], using unions of UMLS semantic types to define allowed values for each query concept. We have used the queries in the system of Berrios and colleagues [9-11] for semi-automated indexing and information retrieval for full-text information resources. During

indexing, a domain expert identifies sentences or paragraphs as answering one or more of the generic queries, and the system automatically proposes values for each query concept. The indexer then specifies the appropriate values. During retrieval, a user selects a generic query appropriate to their question and provides the noun concepts for their search. The search engine then directs the user to the appropriate items of information by matching the user's query to the full-text indexing [12].

Materials and Methods

We collected one hundred and ten consecutive questions asked by physicians in the course of patient care. To obtain information needs from a variety of settings, questions were collected in outpatient, inpatient, and critical-care academic internal medicine. Questions were eligible if they could not be answered by any of the physicians participating at the point of care, resulting in an information need that could be addressed to a clinical knowledge base (regardless of whether a search was actually undertaken). Questions were then grouped by the type of information need expressed as judged by a domain expert in clinical medicine. The resulting categories defined the range of possibilities that the generic queries would encompass.

We then defined a set of generic concepts for the search terms in each query, and thus the allowed values for each principal concept. We defined each generic concept as a union of semantic types from the Unified Medical Language System's semantic network, using the Protégé ontology modeling software [13]. We then combined the defined set of query types with appropriate unions of these dynamic concepts, producing a set of generic queries.

To test the completeness and specificity of the query model, we implemented the query model in the system of Berrios and colleagues to index the full text of sample chapters from widely-used reference texts in general [14] and subspecialty [15] medicine. Indexing was at the paragraph or sentence level, and the indexer was allowed to judge which sentences were sufficiently significant to merit indexing. Indexers tested the query set's completeness (whether a query could be found which matched each unit of information) and the difficulty of choosing the appropriate query.

Results

We produced a query model consisting of thirteen generic queries for clinical information retrieval [Figure 1] that was sufficient to encompass the one hundred and ten real-world clinical questions. We found that four generic concepts, "Manifestation", "Investigation", "Pathology", and "Therapy", were sufficient to express the required

relationships within our query model. We then enumerated the UMLS semantic types which in the judgement of a domain expert in clinical medicine were sufficient to define the generic concepts [Figure 2].

Our preliminary evaluation of the query set has been experiential and qualitative. When used by a group of clinicians uninvolved with the development of the query set to index a sample chapter from a standard medical reference text [14], the queries were sufficiently descriptive to present minimal difficulty in selecting the appropriate query for indexing. Multiple indexing of complex sentences was occasionally necessary. The query set as a whole was sufficiently complete in its coverage of the clinical information to index the large majority of sentences.

When used to index a sample chapter from a subspecialty medical reference text [15], the query set again appeared sufficiently expressive and comprehensive to index the large majority of the material. In a small number of specialized situations, such as the operational details of a procedure, a query could still be found to index the information but with a less complete fit.

In three cases (items 4, 6, and 7 in Figure 1), we defined synonym queries for each parent query (items 4.1, 6.1 and 7.1, respectively). A synonym query is derived by substituting the value "any or none" for one or more concepts (i.e. using existential qualification for the concept.) Indexers use only the parent queries, and during retrieval the synonym queries match their corresponding parents. We modeled these three synonym queries because they were considered among the most common likely to be asked by a clinician user, and would represent an important addition to the search interface.

Discussion

The key trade-off that emerged in our development was between the specificity of each individual query and the comprehensiveness of the entire query set. As each query is made more specific – that is, encompassing a smaller number of potential clinical questions – the query set needs a larger number of queries to remain comprehensive. Highly specific generic queries are likely to improve search precision through more focused searches. However, a very large query set makes the task of selecting the appropriate generic query for indexing or searching increasingly difficult, and may decrease search recall by making it more likely the searcher and indexer will disagree as to which query a unit of indexed information relates. The determination of appropriate query granularity remains a judgement between specificity

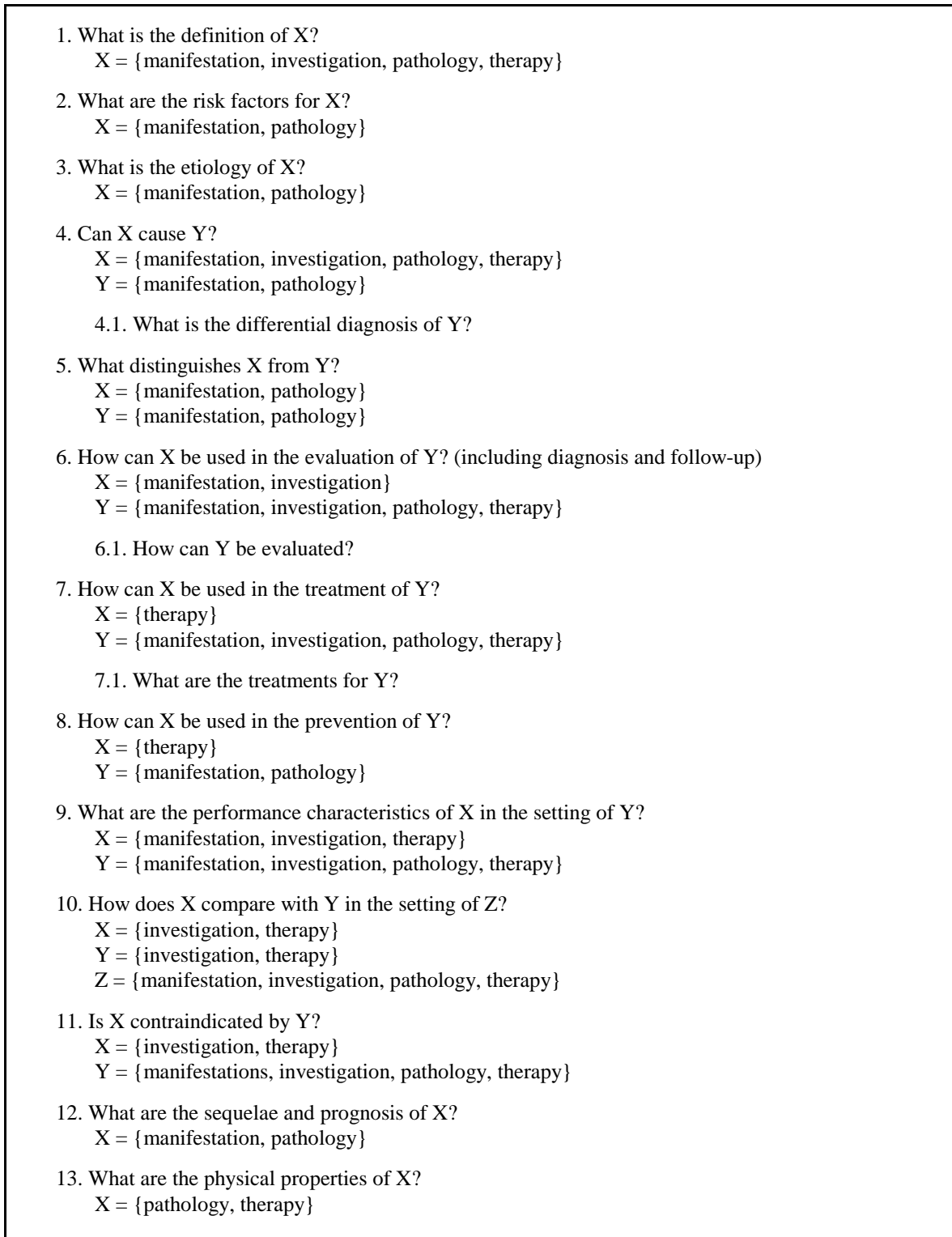


Figure 1 – Generic query model for clinical information retrieval

and query number, and will vary as the query set is adapted to specific sub-domains or applications.

To produce a query set with a tractable number of queries suitable to an indexing task, we have favored as few queries as is consistent with comprehensiveness and mutual exclusivity. In contrast to the prior work of others, we

attained this compactness in part by using broad unions of semantic types to define the query concepts. Additionally, by using broadly defined relationships between concepts, we maintained the independence of each query from any specific sub-domain of clinical

<p><u>Manifestation</u> Finding Laboratory or Test Result Sign or Symptom Clinical Attribute Anatomical Abnormality Acquired Abnormality Congenital Abnormality</p> <p><u>Therapy</u> Clinical Drug Medical Device Biologically Active Substance Neuroreactive Substance or Biogenic Amine Hormone Enzyme Vitamin Immunologic Factor Receptor Biomedical or Dental Material Pharmacologic Substance Antibiotic Element, Ion, or Isotope Food Behavior Social Behavior Individual Behavior Therapeutic or Preventive Procedure</p>	<p><u>Investigation</u> Diagnostic Procedure Laboratory Procedure</p> <p><u>Pathology</u> Anatomical Abnormality Acquired Abnormality Congenital Abnormality Gene or Genome Invertebrate Bacterium Fungus Virus Rickettsia or Chlamydia Hazardous or Poisonous Substance Behavior Social Behavior Individual Behavior Injury or Poisoning Pathologic Function Cell or Molecular Dysfunction Disease or Syndrome Mental or Behavioral Dysfunction Neoplastic Process Experimental Model of Disease</p>
--	---

Figure 2 – Generic noun concepts, defined as unions of UMLS semantic types

medicine. We nonetheless found the generalized queries were effective for indexing full-text medical references, although inevitably some questions will remain outside any explicitly enumerated model of queries. By limiting our query set to questions asked in the context of patient care decision making, we have not modeled queries regarding epidemiology, statistics, medical economics, clinical information retrieval itself, or other sub-fields closely related to clinical medicine. We hypothesize that our query set can itself be used as a basis for more specialized query sets, with more narrowly defined queries and more complex semantic relationships. For example, it could be tailored to individual medical specialties, specific patient care settings, or for differently trained user groups. We are presently developing software tools to facilitate these adaptations. Expanded use of synonym queries may be an important method for increased search specificity by narrowing the focus of search queries and the scope of the noun concepts, without complicating the indexing task.

We defined four superclasses of UMLS semantic types to describe the allowed values for each term in the generic queries. By using the UMLS semantic network we benefit from its continuous refinement and expansion, and can use automated text processing systems like that developed by Berrios and colleagues [9-11] to partially automate the

selection of indexing terms. In order to allow all possible correct values for each term, it was necessary to include a large number of UMLS semantic types. Because of the generality of the UMLS semantic types, some Metathesaurus concepts will be inappropriately included as allowed indexing terms. For example, to include “oxygen” as a type of “Therapy,” it was necessary to include “Element, Ion, or Isotope,” the semantic type of “oxygen” but also of many other concepts that are inappropriate to the type “Therapy.” These semantic type superclasses will allow us to adapt the model to other controlled vocabularies that use semantic typing. Additionally, by modifying the ontology of semantic type superclasses, the query set can be adapted to specific sub-domains or applications. For example, a subset of the types in the superclass “Pathology” could define a superclass “Pathogenic Organisms,” subsuming the appropriate UMLS semantic types. “Pathogenic Organisms” could then be used in a specialization of the query set appropriate to the Infectious Diseases domain.

We are additionally expanding the available specificity for indexing and searching by defining sets of “restriction terms.” These are terms that are not themselves part of a query statement, but restrict the context in which the query applies. For example, “Children” or “Pregnant Patients” can

restrict a query according to patient category, “Emergency Medical Technician” or “Physician” by audience, or “Clinic”, “Hospital”, or “Primitive Setting” by the circumstances of care. Many of these restriction terms and their classification are not modeled in the UMLS semantic network. Restriction sets can be created, shared, and adapted across specialized query sets according to the needs for which they are developed.

Conclusion

We have empirically developed a set of generic queries for clinical information retrieval which preliminary studies suggest is sufficiently comprehensive for full-text indexing and retrieval. The queries use unions of UMLS semantic types to define the allowed values for each dynamic concept, providing for semi-automated selection of indexing terms and for continuous enhancement of the available vocabulary. Future work will include further testing on a variety of electronic clinical information databases, the specialization of the query set according to clinical domain or audience, and the adaptation of the local semantic types to other controlled vocabularies.

Acknowledgements

Dr. Berrios is supported by the Veterans’ Affairs Office of Academic Affairs, Health Services Research and Development Service, and the Office of the Chief Information Officer. Mr. Shah was supported by a scholarship from Harcourt Health Sciences. We thank the National Library of Medicine for experimental access to the UMLS Metathesaurus and to Apelon, Inc. for access to their Metaphrase tools.

References

- [1] Hersh WR, Greenes RA. SAPHIRE – an information retrieval system featuring concept matching, automatic indexing, probabilistic retrieval, and hierarchical relationships. *Comput Biomed Res* 1990; 23:410-425.
- [2] Cimino JJ. Linking patient information systems to bibliographic resources. *Methods Inf Med* 1996; 35:122-126.
- [3] Miller PL, Barwick KW, Morrow JS, Powsner SM, Riely CA. Semantic relationships and medical bibliographic retrieval: a preliminary assessment. *Comput Biomed Res* 1988 Feb; 21(1): 64-77.
- [4] Cimino JJ, Johnson SB, Aguirre A, Roderer N, Clayton PD. The Medline Button. *Proc Annu Symp Comput Appl Med Care* 1992; 81-5.
- [5] Cimino C, Barnett GO, Blewett DR, Hassan LJ, Grundmeier R, Merz R, Kahn JA, Gnassi JA. Interactive Query Workstation: a demonstration of the practical use of the UMLS knowledge sources. *Proc Annu Symp Comput Appl Med Care* 1992; 823-824.
- [6] Joubert M, Robert JJ, Miton F, Fieschi M. The project ARIANE: conceptual queries to information databases. *Proc AMIA Symp* 1996; 378-382.
- [7] Cimino JJ, Elhanan G, Zeng Q. Supporting infobuttons with terminological knowledge. *Proc AMIA Symp* 1997;528-532.
- [8] Ely JW, Osheroff JA, Ebell MH, Bergus GR, Levy BT, Chambliss ML, Evans ER. Analysis of questions asked by family doctors regarding patient care. *BMJ* 1999 Aug 7; 319:358-61.
- [9] Dugan JM, Berrios DC, Liu X, Kim DK, Kaizer H, Fagan LM. Automation and integration of components for generalized semantic markup of electronic medical texts. *Proc AMIA Symp* 1999; 736-40.
- [10] Berrios DC, Kehler A, Fagan LM. Knowledge requirements for automated inference of medical textbook markup. *Proc AMIA Symp* 1999; 676-80.
- [11] Berrios DC. Automated Indexing for Full Text Information Retrieval. *Proc AMIA Symp* 2000; 71-5.
- [12] Kim DK, Fagan LM, Jones KT, Berrios DC, Yu VL. MYCIN II: design and implementation of a therapy reference with complex content-based indexing. *Proc AMIA Symp* 1998; 175-9.
- [13] Musen MA. Domain ontologies in software engineering: use of Protégé with the EON architecture. *Methods Inf Med* 1998; 37:540-550.
- [14] Goldman L, Bennett JC (eds.). *Cecil Textbook of Medicine* 21st ed. WB Saunders: Philadelphia; 1999.
- [15] Braunwald E, Zipes D, Libby P (eds.). *Heart Disease: A Textbook of Cardiovascular Medicine* 5th ed. WB Saunders: Philadelphia; 1996.

Address for Correspondence

Russell J. Cucina MD, S-101 Stanford Hospital, Stanford, California, 94305 ; rjcucina@stanford.edu