Finding the Findings: Identification of Findings in Medical Literature Using Restricted Natural Language Processing

Charles A. Sneiderman, MD, PhD, Thomas C. Rindflesch, PhD, Alan R. Aronson, PhD
National Library of Medicine, Bethesda, MD 20894

The ability to search the biomedical literature based on findings would provide enhanced access to information. We describe a computer program called FINDX which relies on the UMLS® Metathesuaurus® and restricted natural language processing to identify findings in free text. Such identification can serve as a filtering mechanism while selecting relevant papers. After discussing the salient characteristics of findings on which FINDX depends, we report on the results of an experiment in which we tested the program on a set of MEDLINE® abstracts pertaining to the diagnosis of Parkinson Disease.

INTRODUCTION

The health professions are under increasing pressure to utilize the most efficacious and least costly techniques of diagnosis and treatment [1]. Diagnosis depends on findings (i.e. phenomena or entities that are "observations" rather than "conclusions"). Critical appraisal of the significance of findings is an integral part of evidence-based medicine [2], and review of current published medical literature has become the standard of care [3].

Although the Medical Subject Headings used in indexing by the National Library of Medicine contain many "finding" terms, not all of the concepts in the title or abstract of MEDLINE citations can be indexed [4], [5]. Therefore an approach to retrieving medical literature in which findings are described by a restricted natural language approach might be useful in these tasks. We have described a rule-based computer program called FINDX which identifies findings in clinical patient records [6]; this paper will describe an attempt to identify findings in MEDLINE citations of a common disorder, Parkinson disease.

Several approaches to representing the information in medical text have recently been reported ([7], [8], and [9], for example). In addition to addressing the representation of medical knowledge several recent studies explore approaches to discovering this knowledge through natural language processing (NLP) techniques (e.g. [10]). (Pietrzyk [11] provides an overview of recent work in NLP for biomedical text.) Several papers report favorably on using NLP techniques for identifying and representing findings in a particular medical subdomain. See, for example, [12] for asthma discharge summaries; [13] for urology reports; [14], [15], and [16] for chest x-ray reports.

We view our efforts as being complementary to this research, which is concerned (in part) with a detailed representation of the semantic structure of findings. Our goal is to identity those linguistic units which contain a finding, without determining the semantic structure of the finding. This processing might contribute to the efficiency of more detailed semantic analysis by eliminating those sentences which do not contain a finding before the more costly analysis is attempted. The focus of this paper is the further claim that the preliminary processing we propose may have immediate application for specialized tasks like information retrieval.

Before discussing findings in the biomedical literature, we briefly discuss the general characteristics of findings in medical records and review our approach to identifying them as described in [6]. Clinical observations have been classified as subjective or objective in the problem-oriented medical record format, or as derived from interview, physical, laboratory, imaging, or surgical examination in the traditional source-oriented medical record format [17]. However, the language of an observation may be identical regardless of its source; for example, both physical and chest x-ray reports may contain the phrase *lungs are clear*, and both patients and physicians may observe neck mass. Also problematic to a natural language approach to identifying findings is the continuum in which a concept like abdominal pain can be a symptom, an intermediate hypothesis, or a diagnosis, as pathologic and diagnostic processes evolve over time.

FINDX is based on an information science approach to findings as constituting an attribute with implied value (*pain, ascites*) or an attribute with expressed value (*heart rate 110, elevated liver function tests*). The program relies crucially on the UMLS Metathesaurus (6th Experimental Edition, April 1995), and in particular on the UMLS semantic types, in order to

determine finding attributes. FINDX values can be either a number or one of a list of modifiers derived from SNOMED International. This list was compiled from the general adjectival modifiers and from adjectives in the Function and Morphologic axes. We propose four rules consistent with this approach, which identify finding attributes and values generally in biomedical free text.

In medical records, mention of a diagnostic procedure without a value can indicate an order for a test, and not the result. For example, (1) may occur in a record before the results are reported as (2).

(1) We suggest arterial blood gas pre-operatively

(2) arterial blood gas 7.41/42/43/27

In order to identify the latter as a finding, but not the former, we formulate the rules so that a value is required for successful application.

An initial review of findings in biomedical literature suggested that the rules as articulated for medical records would have to be modified in order to effectively identify findings in MEDLINE abstracts. The literature sometimes implicates or hypothesizes results or relationships as in (3).

(3) Mean systolic and diastolic blood pressure differed significantly among the three groups.

In (3) the actual value of the blood pressure is not given. The implication is that there is a result, and thus this is a finding with respect to the disorder being discussed.

In order to accommodate findings in the literature, we conducted an experiment in which the original FINDX rules were relaxed slightly, in that they were not required to have values. However, the results with regard to precision and recall were not significantly different from those obtained by applying the rules configured as they had been for the medical records, that is, by requiring these rules to have values. It would seem that examples like those in (4) predominate.

(4) There has been some debate about abnormalities in visual evoked potentials (VEP) in Parkinson's Disease (PD).

HM-PAO SPECT showed marked perfusion asymmetry in parietal cortical regions...

Both examples contain values: *abnormalities* in the first and *marked perfusion asymmetry* in the second.

In the majority of instances the structure of findings in the medical literature very closely parallels the structure of findings in medical records. The same rules can effectively be used in both types of text for the identification of findings.

METHODS

As noted above we do not at this time attempt to impose a semantic representation on the findings we identify. Rather, we seek broad application of our method; we have so far tested FINDX on a variety of medical records, and in this paper report on its application to the biomedical research literature.

A search of the current MEDLINE database was performed on December 4, 1995 to retrieve English-language citations for articles indexed principally by MeSH heading Parkinson Disease combined with sub-heading Diagnosis. Sixty-six citations were retrieved. The titles and abstracts of the citations contained about 10,000 words in 497 sentences, which were marked by the first author, a physician, if he judged them to contain findings. 237 sentences were determined to contain findings.

This text was then submitted to processing with the FINDX program. Text to be analyzed is first subjected to a preprocessor which attempts to identify sentences and other significant linguistic units, such as complex noun phrases and sentence fragments. Syntactic processing then applies to each linguistic unit identified by the preprocessor. The main goal of this processing is to identify simple noun phrases. Syntactic analysis is supported by the SPECIALIST Lexicon [18] and the Xerox part-of-speech tagger [19]. The result is an underspecified syntactic structure in which simple noun phrases are identified through efficient processing that does not have to be tuned for specific domain areas. For example, input text (5) is delimited into phrases as shown in (6).

(5) Magnetic resonance imaging revealed irregular patchy areas of increased signal intensity.

(6) [magnetic resonance imaging] [revealed] [irregular patchy areas] [of increased signal intensity]

Once the underspecified syntactic structure has been determined, FINDX maps each noun phrase to concepts in the UMLS Metathesaurus using the MetaMap program [20]. In the current example, such processing determines that the text *magnetic resononance imaging* maps to the Metathesaurus concept "Magnetic Resonance Imaging" with semantic type 'Diagnostic Procedure'.

The final step in FINDX processing is to apply the finding rules, which take advantage of information

gathered in the previous steps, in particular the UMLS semantic types. FINDX currently has four rules: The Anatomy Rule, The Physiologic Function Rule, The Test Result Rule, and The Sign or Symptom Rule. In the following discussion of these rules we mention some general considerations in their formulation and then give examples of their application to text from patient records and from the MEDLINE abstracts which consitute our test set for this study.

The Anatomy Rule (7) is formulated to identify those findings which constitute a comment on some characteristic of an anatomical entity.

(7) The Anatomy Rule

Attribute: UMLS semantic types: 'Acquired Abnormality', 'Body Location or Region', 'Body Part, Organ, or Organ Component', 'Body Space or Junction', 'Body System', 'Congenital Abnormality', 'Embryonic Structure', 'Fully Formed Anatomical Structure', 'Tissue', 'Cell', 'Cell Component'.

Value: SNOMED adjective.

Examples to which this rule applies typically include findings from the physical examination, such as: *chest clear to auscultation* or *ears negative.* However, the rule can also apply to image findings (*chest X-ray showed normal heart size*) or tissue findings (*serosa is pink*). In the literature, the Anatomy Rule applies to a sentence such as (8).

(8) Eight demented cases had absent neocortical neurofibrillary tangles.

The concept "Neurofibrillary Tangles" has semantic type 'Cell Component' and thus satisfies the attribute part of the rule. The value part of the rule is satisfied by the word *absent.*

The Physiologic Function Rule (9) refers to an attribute covered by the three semantic types noted; the value may either be quantitative or qualitative.

(9) The Physiologic Function Rule

Attribute: UMLS semantic types: 'Physiologic Function', 'Organism Function', 'Organ or Tissue Function'.

Value: numeric or SNOMED adjective.

Examples from patient records include *respiratory status stable;* where the concept "Respiration" has semantic type 'Physiologic Function'; *appetite normal* ('Organism Function'); and *blood pressure 130/90* ('Organ or Tissue Function'). The example in (10), from a MEDLINE abstract, satisfies the rule in that "Tendon Reflexes" has semantic type 'Organ or Tissue Function' and *increased* is a SNOMED adjective.

(10) Increased tendon reflexes associated or not with frank pyramidal signs...are highly suggestive of the disease.

The Test Result Rule (11) identifies findings which are results of diagnostic tests.

(11) The Test Result Rule

Attribute: UMLS semantic types: 'Diagnostic Procedure', 'Laboratory Procedure', 'Laboratory or Test Result'.

Value: numeric or SNOMED adjective.

Examples from medical records are *echocardiography preliminary report showed small posterior effusion* ("Echocardiography" has semantic type 'Diagnostic Procedure'); *prothrombin time normal* ("Prothrombin Time" has semantic type 'Laboratory Procedure'); *no weight gain* ("Weight Gain" has semantic type 'Laboratory or Test Result' in addition to 'Organism Function'). In (12), which is from an abstract, "PET Scan" has semantic type 'Diagnostic Procedure' and *normal* satisfies the value part of the rule.

(12) Fluorodopa F 18 (F-dopa) positron emission tomographic scanning yielded normal findings in three patients.

UMLS has three semantic types directly relating to findings, namely 'Finding', 'Sign or Symptom', and 'Pathologic Function'. The Sign or Symptom rule (13) applies whenever text maps to a UMLS concept having any of these semantic types.

(13) The Sign or Symptom Rule

Attribute: UMLS Semantic types: 'Finding', 'Pathologic Function', 'Sign or Symptom'.

Value: No value specified.

The "findings-diagnosis continuum" is represented in the UMLS Metathesaurus by concepts having multiple semantic types. A number of Metathesaurus concepts ("Angina Pectoris" and "Seizures," for example) have semantic type 'Disease or Syndrome' in addition to 'Sign or Symptom'. Metathesaurus concepts which have semantic types 'Disease or Syndrome' and 'Finding' include "Cerebral Infarction" and "Ventricular Tachyarrhythmia."

If text maps to a Metathesaurus concept which has semantic type 'Finding' or 'Sign or Symptom', we always consider it as a finding, regardless of the other

semantic types it may also have. Given the ambiguity which occurs out of context between a finding and a diagnosis, we believe that it is better to let the searcher determine the concept's use. In an example from the literature, *Autonomic failure, depression and anxiety in Parkinson's disease*, the Metathesaurus concepts "Failure," "Depression," and "Anxiety" have semantic type 'Finding.'

In (14) the concept "Cerebral Vascular Lesion" has semantic type 'Pathologic Function', and the Sign or Symptom Rule would thus identify this as a finding.

(14) Multiple cerebral vascular lesions on MRI correlated significantly and independently with the extent of the PVH.

The semantic type 'Pathologic Function' includes normal responses to negative stimuli as well as pathologic conditions that are less specific than a disease. When we examined the Metathesaurus terms so typed, we found many to be observations. They are often equivalent to physiologic functions with value, e.g. "Achlorhydria," or anatomic entity with value, e.g. "Pyocolpos," or precoordinated terms such as "Atrial Fibrillation" and "Neoplasm Seeding."

RESULTS

When the FINDX methodology was applied to our test set of MEDLINE abstracts, the program was able to identify 194 of the 237 marked findings. Since 43 marked findings were missed by the program and 103 sentences were incorrectly identified as containing a finding, recall is 82% and precision is 65%. As a first step toward improving these figures, in the following section we discuss the error types so far identified.

DISCUSSION

Marked findings missed by the program (false negatives) fall into several categories. Acronyms and abbreviations are a major concern in any NLP project. Although they do not appear as frequently in biomedical literature as in patient records, it is common for the literature to abbreviate a recurring phrase representing one of the principal variables described, as, for example, when *visual evoked potential* is abbreviated to *VEP*. This particular acronyn does not occur in the knowledge sources available to FINDX and thus utterances containing strings like *abnormal VEPs* were not identified as findings even though "Evoked Potentials, Visual" is a Metathesaurus concept with semantic types 'Diagnostic Procedure' and 'Organ or Tissue Function'.

As mentioned above, we required findings to have overt values, although a few literature findings, such as the example in (3) above, do not meet this requirement. Such findings were missed by the program.

It is also common for scientific literature to describe novel associations for current concepts and to use new terms or combinations of existing terms to describe new phenomena. FINDX is dependent upon the knowledge contained in the UMLS Metathesaurus, which is based on terminology old enough to have been included in a controlled vocabulary. Thus, (15) was not identified as containing a finding.

(15) Brain beta 2-microglobulin levels are elevated in Parkinson's disease.

Although the concept "beta 2-Microglobulin" occurs with semantic type 'Amino Acid, Peptide, or Protein', the program did not have enough information to determine that a laboratory test result is being reported.

Although the UMLS Metathesaurus has broad coverage of biomedical concepts it is unrealistic to expect it to be complete. The concept *P100 latencies* is not currently included and hence FINDX did not recognize (16) as a finding.

(16) The PD patients treated with levodopa had significantly longer P100 latencies than the other PD patients.

In a few instances the value of the finding is expressed in terms too complex for our processing to recognize. For example, in (17), FINDX did not recognize that *could be held tightly to the side* was a comment on the arm.

(17) However, the arm could be held tightly to the side.

Sentences incorrectly identified as containing a finding (false positives) are due to the inappropriate application of a FINDX rule. In general this occurs because our methodology depends on an underspecified syntactic analysis. In (18), FINDX wrongly interprets the SNOMED modifer *limited* as being a comment on the anatomic concept "Rectum" rather than as a comment on routes of administration.

(18) Local allergic effects have limited the use of other routes of administration, such as intranasal, sublingual, and rectal routes.

A more detailed syntactic analysis would be required in order to eliminate this as a possible interpretation. In the text *high field MRI* the concept "MRI" has semantic type 'Diagnostic Procedure', but our syntactic analysis is not sophisticated enough to support the correct interpretation in which *high* modifies that procedure. Rather, we incorrectly interpret *high* as being the value obtained from an MRI report.

CONCLUSION

Review of the medical literature for clinical decision making is becoming increasingly common. Such projects as the Cochrane collaboration (an attempt to develop a body of meta-analyses of clinical trials data in all the clinical problems where such data exists) require selective assembly of massive amounts of published information [21]. The purpose of the periodical literature of science is to report new observations or findings. We believe that the necessity to inferentially determine "new knowledge" will always exceed the capacity of the best indexing systems based on controlled vocabularies derived from current biomedical terminology. Natural language processing efforts such as the one described above may be fruitful in retrieving information from patient records and medical literature.

References

[1] Lohr KN. Guidelines to clinical practice: What they are and why they count. *J Law Med Ethics*, 1995; 23:49-56.

[2] Jaeschke R, Guyatt GH, Sackett DL. Users' Guide to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. *JAMA*, 1994; 271(9):703-7.

[3] Oxman AD, Sackett DL, Guyatt GH. Users' guides to the medical literature. I. How to get started. The Evidence-based Medicine Working Group. *JAMA*, 1993; 270(17):2093-5.

[4] Sneiderman CA. Keeping up with the literature of family practice: a bibliometric approach, *Fam Pract Res J*, 1983; 3:17-21.

[5] Sneiderman CA. Medical terminology activities at the National Library of Medicine. *J Clin Comput*, 1983; 12:36-9.

[6] Sneiderman CA, Rindflesch TC, Aronson AR, Browne AC. Extracting physical findings from free-text patient records. *Abstracts of AMIA Spring Congress*, 1995:49.

[7] Friedman C, Huff SM, Hersh WR, Pattison-Gordon E, Cimino JJ. The Canon Group's effort: Working toward a merged model. *JAMIA*, 1995; 2:4-18.

[8] Rector AL, Glowinski AJ, Nowlan WA, Rossi-Mori A. Medical-concept models and medical records: an approach based on GALEN and PEN&PAD. *JAMIA*, 1995; 2(1):19-35.

[9] Pattison-Gordon E, Greenes RA. An empirical investigation into the conceptual structure of chest radiographic findings. *Proc Annu Symp Comput Appl Med Care*, 1994:257-261.

[10] Rassinoux AM, Wagner JC, Lovis C, Baud RH, Rector A, Scherrer JR. Analysis of medical texts based on a sound medical model. *Proc Annu Symp Comput Appl Med Care*, 1995:27-31.

[11] Pietrzyk PM. Free text analysis. *Int J Biomed Comput*, 1995; 39(1):139-44.

[12] Sager N, Lyman M, Bucknall C, Nhan N, Tick LJ. Natural language processing and the representation of clinical data. *JAMIA*, 1994; 1(2):142-60.

[13] el-Gamal SS, Esmail MM. Understanding clinical narrative text. *Med Inf (Lond)*, 1995; 20(2):161-73.

[14] Haug PJ, Koehler S, Lau LM, Wang P, Rocha R, Huff SM. Experience with a mixed semantic/syntactic parser. *Proc Annu Symp Comput Appl Med Care*, 1995:284-8.

[15] Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. *JAMIA*, 1994; 1(2):161-74.

[16] Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. Unlocking clinical data from narrative reports: a study of natural language processing. *Ann Intern Med*, 1995; 122(9):681-8.

[17] Weed LL, Medical records that guide and teach. *N Engl J Med*, 1968; 278(11):593-600; 278(12):652-7.

[18] McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. *Proc Annu Symp Comput Appl Med Care*, 1994:235-9.

[19] Cutting D, Kupiec J, Pedersen J, Sibun P. A practical part-of speech tagger. *Proceedings of the Third Conference on Applied Natural Language Processing*, 1992:133-140.

[20] Aronson AR; Rindflesch TC; Browne AC. Exploiting a large thesaurus for information retrieval. *Proceedings of RIAO 94*, 1994:197-216.

[21] Bero L; Rennie D, The Cochrane Collaboration. Preparing, maintaining, and disseminating systematic reviews of the effects of health care. *JAMA*, 1995; 274(24):1935-8.