Methods of Information in Medicine
© F. K. Schattauer Verlagsgesellschaft mbH (1993)

J. J. Cimino[1], G. O. Barnett[2]

# Automatic Knowledge Acquisition from MEDLINE

[1]Center for Medical Informatics,
Columbia University, Columbia-Presbyterian Medical Center,
New York, NY, USA
[2]Laboratory of Computer Science,
Harvard Medical School,
Massachusetts General Hospital,
Boston, Mass., USA

**Abstract:** Construction of medical knowledge bases for use in expert systems is an arduous task. We propose a procedure for obtaining medical knowledge via automated analysis of citations found in the National Library of Medicine's MEDLINE® database. In this method, simple pattern of keywords and subheading co-occurrences are detected in the keyword descriptor portion of the citations. Each pattern corresponds to a fact, expressed as a semantic relationship between medical concepts. We have constructed a set of 504 pattern-matching rules and applied it to a set of 673 MEDLINE® citations to produce 2,795 such facts. The results are presented of an analysis of the syntactic and semantic features of these facts to understand the kinds of knowledge than can be obtained through our method and speculate on the potential uses and pitfalls for knowledge of this type.

***Key-Words:*** Computer Reasoning, Knowledge Acquisition, Knowledge Bases, MEDLINE

## 1. Introduction

The construction of medical knowledge bases such as those used in expert systems is arduous, in part because extensive reviews of medical literature are required [1, 2]. Until now, the process of literature review for the accumulation of medical knowledge has been performed by medical experts or their assistants. An automated assistant which could review the literature, propose medical knowledge for inclusion in a knowledge base, and support the proposal with references to the literature could be enormously useful to knowledge engineers. The automated abstraction of medical literature has proved difficult due to the complexity of natural language, compounded by the fact that the world medical literature is written in at least 43 different languages. There has been some success with automatically classifying medical text [3], but the automatic extraction of meaning from such text has remained elusive. For example, the originators of Roundsman, an expert system which draws its knowledge directly from the medical literature, were still required to produce their knowledge base through the manual abstraction of journal articles to create highly structured, declarative representations of clinical studies [4].

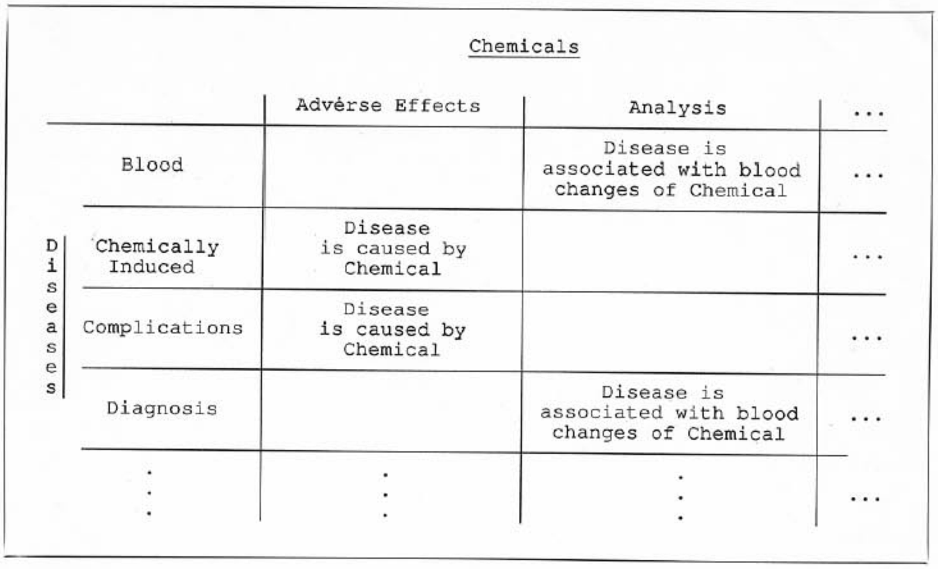
The National Library of Medicine (NLM) uses highly trained indexers to review journal articles and create citations which describe the content of those articles [5]. Unlike the literature from which they have been abstracted, these citations use a well-defined structure and a controlled vocabulary. The citations, which form the MEDLINE® database, make use of the *Medical Subject Headings* (MeSH®) to convey the topics discussed in the cited literature [6]. Additional information regarding the context of these topics is portrayed through the use of a set of term qualifiers, called *Subheadings,* that are applied to MeSH headings.

Although there are no explicit relationships among MeSH terms included in a citation, relationships among the terms may often be inferred based on the contexts in which the terms appear (obtained by examining the Subheadings applied to each term). Furthermore, these relationships are semantic in nature and are, therefore, a form of medical knowledge. For example, if a citation indicates that a particular journal article discusses the cause of Disease X and the adverse effects of Drug Y, it is reasonable to propose that "Drug Y *causes* Disease X". Anyone wishing to verify that statement needs merely to refer to the cited article to determine the validity of this (semantic) relationship between the drug and the disease.

Powsner and coworkers have examined the semantic relationships between MeSH terms in MEDLINE citations to test their usefulness in performing literature searches [7]. We propose that, through the use of simple pattern-matching rules, it is possible to automatically generate semantic relations between medical concepts for use in medical knowledge bases. The process by which these relationships are generated has been described [8]; this paper examines the types of relationships that can be generated, and discusses ways in which this approach can assist in building medical knowledge bases.

## 2. Methods

Pattern-matching rules are represented as cells in simple matrices called Rule Tables, where each cell corresponds to a possible rule and, when the rule exists, contains a potential fact which expresses some semantic relationship between medical concepts. The Tables were constructed using one set of MEDLINE citations and evaluated using a second set of citations. The semantic relationships generated in this process were then examined to determine general characteristics of the kinds of knowledge that can be extracted from the citations. The methods used and results obtained in Rule Table construction and validation are reported elsewhere [8] and are recounted briefly here to clarify discussion of the analysis of the resulting semantic relations.



| | | Chemicals | | |
|---|---|---|---|---|
| | | Advérse Effects | Analysis | ... |
| Diseases | Blood | | Disease is associated with blood changes of Chemical | ... |
| | Chemically Induced | Disease is caused by Chemical | | ... |
| | Complications | Disease is caused by Chemical | | ... |
| | Diagnosis | | Disease is associated with blood changes of Chemical | ... |
| | ⋮ | ⋮ | ⋮ | ... |

**Figure 1** Portion of a rule Table concerning MeSH Diseases and Chemicals. The top row shows two of the 26 Subheadings which are used with Chemicals and the left-most column shows four of the 36 Subheadings used with Diseases. When a MEDLINE Citation contains a Disease term and a Chemical term as descriptors, and these terms are associated with Subheadings, the algorithm looks in the appropriate cell of the table for the relationship between the two terms. Empty cells indicate that either no consistent relationship can be found or that sufficient citations with the specific combination have yet to be examined to detect possible relationships.

### *Rule Table Design*

Powsner et al. [7] have established a set of "key topics" (such as "micronodular versus macronodular cirrhosis in hepatocellular carcinoma"), identified journal articles that were relevant to each topic, and then examined the MEDLINE citations for consistent patterns among the MeSH Headings and Subheadings. These patterns were used to formulate search strategies for citations relevant to the topic. We describe a reverse approach, toward achieving an opposite goal: patterns are identified in MEDLINE citations which are then correlated with the topics presented in the article. The patterns consist of co-occurrences of classes of MeSH Headings and specific Subheadings (such as "Disease/ETIOLOGY and Disease/COMPLICATIONS"). The topics consist of semantic relationships between medical concepts (such as "myocardial infarction is caused by coronary artery disease"). When a recurring Heading/Subheading pattern is found in additional citations, a relationship is proposed between the classes in the pattern.

Simple pattern-matching "If-Then" rules are then constructed from the MEDLINE citations. Each rule examines a pair of MeSH terms belonging to any of five particular classes (or MeSH "Trees"): *Anatomical; Cytological and Embryological Terms* (referred to in this paper as Anatomical Sites); *Biological Sciences* (Biological Processes); *Chemicals and Drugs* (Chemicals); *Diseases, Symptoms and General Pathology* (Diseases); and *Analytical, Diagnostic and Therapeutic Techniques* (Procedures). The remaining 11 classes of MeSH terms were not examined. The general form of the rule is:

If Class(⟨MeSH A⟩) = ⟨Class 1⟩ AND
⟨MeSH A⟩/⟨Subheading X⟩ AND
Class(⟨MeSH B⟩) = ⟨Class 2⟩ AND
⟨MeSH B⟩/⟨Subheading Y⟩ THEN
⟨MeSH A⟩⟨Relationship Z⟩⟨MeSH B⟩,

which can be read as:

"If a citation contains some MeSH Term A which is of Class 1 AND appears with Subheading X AND the citation also includes some MeSH Term B which is of Class 2 AND appears with Subheading Y THEN propose that A and B are related through the Relationship Z."

An example of a specific rule, in a somewhat simpler format, is:

If
⟨Disease⟩/CHEMICALLY INDUCED
AND
⟨Chemical⟩/ADVERSE EFFECTS
THEN
⟨Disease⟩ is caused by ⟨Chemical⟩.

In general, MeSH terms were considered regardless of their assignment as major or minor descriptors in the citations. Sometimes, a recurring pattern was found but a relationship could only be proposed for some of the examples. Patterns can be modified to require that one or both of the MeSH terms appear as major descriptors in the citations. In some of these cases, we found that when this modification was made, irrelevant citations were filtered out, improving the specificity of the rule. This filter was used on a rule-by-rule basis, rather than on the entire rule set so that sensitivity would not be unduly restricted. A rule thus modified would be of the form:

**Table 1** Rule Tables Sizes

| First Class | Second Class | Total Possible* | Actual Number** |
|---|---|---|---|
| Diseases | Anatomic Sites | 900 | 77 |
| Diseases | Biologic Processes | 612 | 8 |
| Diseases | Chemicals | 936 | 76 |
| Diseases | Diseases | 1296 | 102 |
| Diseases | Procedures | 540 | 21 |
| Procedures | Anatomic Sites | 375 | 85 |
| Procedures | Biologic Processes | 255 | 2 |
| Procedures | Chemicals | 390 | 37 |
| Procedures | Diseases | 540 | 68 |
| Procedures | Procedures | 225 | 28 |

* Based on the product of the number of subheadings allowed for the first and second classes.

** The values given represent the number of cells in each table which contain a proposed relationship. Some 63 additional cells contain markers indicating that no consistent relationship exists. The remaining 5,502 cells are simply empty, indicating that those permutations of Subheading pairs have not yet been examined because their co-occurrence did not appear in the training set.

**Table 2** Some Relationships involving Diseases or Procedures

| Relationship | Type* |
|---|---|
| Disease affects Anatomic Site | I |
| Procedure is performed on Anatomic Site | I |
| Disease is associated with blood changes of Anatomic Site | I |
| Disease is caused by some chemical affecting Anatomic Site | II |
| Disease is caused by some other disease affecting Anatomic Site | II |
| Disease is treated by some procedure affecting Anatomic Site | II |
| Procedure has chemical part which affects Anatomic Site | II |
| Procedure is part of some other procedure on Anatomic Site | II |
| Disease is related to Anatomic Site | IV |
| Disease is caused by some other disease affecting Biologic Process | II |
| Disease is treated by some chemical affecting Biologic Process | II |
| Procedure has chemical part which affects Biologic Process | II |
| Disease is caused by Chemical | I |
| Disease is prevented or controlled by Chemical | I |
| Procedure has part Chemical | I |
| Disease is associated with urine changes of Chemical | I |
| Disease is diagnosed by some procedure with Chemical (as a part) | II |
| Disease is related to Chemical | IV |
| Disease is caused by some other Disease | I |
| Procedure treats Disease | I |
| Procedure diagnoses Disease | I |
| Disease is treated by some chemical causing another Disease | II |
| Disease is caused by some procedure treating another Disease | II |
| Disease changes some blood chemical in some other Disease | II |
| Procedure has chemical part which causes Disease | II |
| Procedure is part of other procedure which diagnoses Disease | II |
| Disease is caused by same chemical as another Disease | III |
| Disease is treated by same chemical as another Disease | III |
| Disease has same urine chemical change as another Disease | III |
| Disease is treated by same procedure as another Disease | III |
| Disease is related to some other Disease | IV |
| Procedure is related to Disease | IV |
| Disease is diagnosed by Procedure | I |
| Disease is prevented or controlled by Procedure | I |
| Disease is caused by some chemical part of Procedure | II |
| Disease is diagnosed by a Procedure which is part of some other procedure | II |
| Procedure is related to some other Procedure | IV |

* Types I, II, III and IV are described in the text. There are no Type III relationships for Procedures.

If
⟨Disease⟩/CHEMICALLY INDUCED
[Major Descriptor]
AND
⟨Chemical⟩/ADVERSE EFFECTS
[Major or Minor Descriptor]
THEN
⟨Disease⟩ is caused by ⟨Chemical⟩.

All of the rules are pairwise comparisons with can be organized into two-dimensional tables, where each row corresponds to a Subheading used with one class of MeSH terms and each column corresponds to a Subheading used with the second class of MeSH terms; there is one table for each comparison of two classes (Diseases and Chemicals, Procedures and Chemicals, Diseases and Diseases, etc.). When a rule is added to the rule base, the hypothesis (i.e., the proposed relationship) is entered into the cell of the table corresponding to the column of one Subheading and the row of the other Subheading. In the example given above, the relationship "Disease is caused by Chemical" would be found in the table relating Diseases and Chemicals in the row for CHEMICALLY INDUCED and the column for ADVERSE EFFECTS (see Figure 1).

### *Rule Table Construction*

A training set of MEDLINE citations was produced by performing 19 MEDLINE searches: one each for two cardiovascular diseases (Myocardial Infarction and Syncope), two cardiovascular procedures (Heart Auscultation and Angiocardiography), and each of fifteen Subheadings used with MeSH diseases (regardless of the disease with which they appeared in the citation). The relationships between disease or procedure terms and other terms in the same citations were then examined. From these examinations, a set of Rule Tables was generated manually.

The rules were applied to the training set of citations and a set of proposed relationships between specific MeSH terms was generated automatically. Each relationship was traceable to the satisfaction of at least one specific rule by one or more specific citations. The citations responsible for each relationship were examined manually by one of us (JJC) to determine if they supported the proposed relationship. Relationships were judged correct if review of the citations (titles or abstracts) revealed some supporting statement and revealed no refuting statement. When examination of the title and abstract was insufficient to verify a relationship, the text of the cited article was examined to determine if a relationship was correct. When relationships were found to be unsupported by their citations, they were judged incorrect. Any rule which suggested an incorrect relationship was labeled invalid and was removed from the Rule Table.

**Table 3** Summary of Relationships found for Aortic Valve Stenosis and Echocardiography

| Relationships | Type* |
|---|---|
| Aortic Valve Stenosis affects 12 Anatomic Sites | I |
| Aortic Valve Stenosis is treated by 2 Chemicals | I |
| Aortic Valve Stenosis is associated with blood changes of 6 Chemicals | I |
| Aortic Valve Stenosis causes 13 Diseases | I |
| Aortic Valve Stenosis is caused by 7 Diseases | I |
| Aortic Valve Stenosis is caused by 3 Procedures | I |
| Aortic Valve Stenosis is diagnosed by 7 Procedures | I |
| Aortic Valve Stenosis is treated by 7 Procedures | I |
| Echocardiography is performed on 32 Anatomic Sites | I |
| Echocardiography has 12 Chemical parts | I |
| Echocardiography diagnoses 79 Diseases | I |
| Aortic Valve Stenosis is caused by chemicals affecting 1 Anatomic Site | II |
| Aortic Valve Stenosis is caused by procedures affecting 1 Anatomic Site | II |
| Aortic Valve Stenosis is diagnosed by procedures with 1 Chemical | II |
| Aortic Valve Stenosis is caused by procedures treating 7 Diseases | II |
| Echocardiography has chemical parts which affect 10 Biologic Processes | II |
| Echocardiography is part of a procedure which diagnoses 3 Diseases | II |
| Aortic Valve Stenosis and 1 other Disease are treated by same chemicals | III |
| Aortic Valve Stenosis and 18 other Diseases are treated by same procedure | III |
| Aortic Valve Stenosis is related to 68 other Diseases | IV |
| Echocardiography is related to 24 Procedures | IV |

* Types I, II, III and IV are described in the text. There are no Type III relationships for Procedures. The complete listing of the counts for each relationship can be found in [8].

### *Rule Table Evaluation*

The evaluation of the rule base was addressed by producing an evaluation set based on searches for citations containing either of the MeSH terms "Echocardiography" (a Procedure) or "Aortic Valve Stenosis" (a Disease). The rule base was applied to the evaluation set and the resulting relationships were compared to the citations from which they were derived.

### *Analysis of Semantic Relationships Among MeSH Terms*

A simple grammar was developed to depict the syntax of proposed relationships. The medical concepts (i. e., MeSH terms) serve as nouns and are represented by upper-case letters **(X, Y, Z,..)**. Semantic relations serve as verb phrases and are signified by lower-case letters **(a, b,..)**. For example, a relationship such as "Chemical causes Disease" would be parsed into the form **X a Y**, where **X** represents "Chemical", **Y** represents "Disease", and **a** represents "causes". A special noun, signified by ***U***, is used to represent a term that is implied by the relationship but whose actual value is unknown (that is, unbound). For example, the relationship "Procedure has some chemical which causes Disease", would be parsed as **X a *U* b Y**, where **X** represents "Procedure", **Y** represents "Disease", **a** represents "has", **b** represents "causes", and ***U*** represents "some chemical". In an actual proposed relationship (such as "angiography has some chemical which causes renal failure"), the procedure and disease would be made explicit, while the chemical would remain unspecified. This grammar is similar to that used in computational linguistics [9].

## 3. Results

### *Rule Table Construction*

The 19 MEDLINE searches (myocardial infarction, syncope, heart auscultation, angiocardiography, and the 15 disease Subheadings) produced a training set of 2,383 citations. Manual examination of the citations resulted in an initial set of Rule Tables containing 567 rules. Application of the initial Rule Tables to the training set produced a list of proposed relationships. Sixty-three rules were found which proposed invalid relationships (based on the citations from which they were derived). In each case, one or more citations were found that produced exceptions to a rule which had appeared valid in one particular citation on initial review.

The offending rules were eliminated from the Rule Tables, leaving 504 rules capable of producing 59 different semantic relationships. Table 1 lists the sizes of the final Rule Tables; Table 2 lists examples of the semantic relations which involve Diseases or Procedures.

### *Rule Table Evaluation*

The MEDLINE search "Echocardiography or Aortic Valve Stenosis" yielded 673 citations for the Evaluation Set. When the Rule Tables were applied to the evaluation set, 2,795 relationships among 286 MeSH terms in the citations were proposed. Three-hundred-sixty-four (364) of these relationships involved the original search topics (173* relationships between

* In the results as originally reported, Echocardiography was listed as having 180 relationships, including the diagnosis of 77 diseases and the cause of 9 diseases [8]. As noted in [8], a transcription error was found in the Rule Table which produced erroneous results. For this report, the results are given with that error corrected; specifically, the 9 diseases previously listed as "caused" by Echocardiography were actually "diagnosed". The combination of these 9 diseases with the original 77 diseases resulted in 2 new diseases (a total of 79 unique diseases) being diagnosed by Echocardiography, for a new total of 173 relationships.

Echocardiography and other terms, 191 relationships between Aortic Valve Stenosis and other terms).

When these 364 relationships were examined, 347 (95%) were felt to represent the relationships that actually existed in the title, abstract or text of the articles whose citations caused the relationships to be proposed. The remaining 17 (5%) were judged to be incorrect representations of the actual relationships in the citations; the 12 rules which caused the proposal of these relationships (2.4% of the 504 rules) were, therefore, deemed to be inaccurate (or, more properly, inconsistent when applied to the new domain). For example, one rule postulated that:

If
⟨Procedure⟩/INSTRUMENTATION
AND
⟨Anatomic Site⟩/ABNORMALITY
THEN
⟨Procedure⟩
is performed on ⟨Anatomic Site⟩.

However, processing the evaluation set produced the proposition that "Echoacardiography is performed on Femur". The responsible citation had the MeSH term-Subheading pattern corresponding to the above rule, but the citation was actually about fat embolism (caused by femur fracture) and its diagnosis by echocardiography. Since the echocardiography was not performed on the femur, the proposed relationship was judged invalid and the responsible rule was regarded as inconsistent.

Table 3 lists examples of the number of MeSH terms found to be related (to Aortic Valve Stenosis and Echocardiography) through each of the different types of relationships after processing the evaluation set. (For a complete list of these numbers, see [8]). Some specific examples from the different categories of relationships are shown in Table 4.

### *Analysis of Semantic Relationships Among MeSH Terms*

Examination of the syntax of the relationships revealed four recurring patterns or types, identified with Roman numerals I–IV. The structure of each of these types is described below.

**Table 4** Examples of Relationships Automatically Extracted from MEDLINE Citations

| Relationship | Type* |
|---|---|
| Aortic Valve Stenosis causes Heart Enlargement | I |
| Aortic Valve Stenosis has blood changes of Neuroregulators | I |
| Aortic Valve Stenosis is treated by Angioplasty, Transluminal | I |
| Echocardiography diagnoses Aortic Valve Stenosis | I |
| Echocardiography prevents or controls Pulmonary Embolism | I |
| Echocardiography has chemical part Isotonic Solutions | I |
| Echocardiography is performed on Heart Atrium | I |
| Aortic Valve Stenosis is diagnosed by procedure with Technetium | II |
| Aortic Valve Stenosis is caused by disease affecting Aortic Valve | II |
| Aortic Valve Stenosis is caused by procedure treating Myxoma | II |
| Echocardiography has chemical part which affects Blood Flow Velocity | II |
| Aortic Valve Stenosis is caused by procedures affecting Aortic Valve | II |
| Aortic Valve Stenosis is caused by procedures treating Myxoma | II |
| Aortic Valve Stenosis and Heart Enlargement are treated by same chemical | III |
| Aortic Valve Stenosis and Mitral Valve Stenosis treated by same procedure | III |
| Aortic Valve Stenosis is related to Angina Pectoris | IV |
| Aortic Valve Stenosis is related to Myocardium | IV |
| Aortic Valve Stenosis is related to Peptide Hydrolases | IV |
| Echocardiography is related to Monitoring, Physiologic | IV |
| Echocardiography is related to Prenatal Diagnosis | IV |

* Some of these relationships will be familiar as common knowledge (such as "Echocardiography diagnoses Aortic Valve Stenosis") while others may seem somewhat peculiar when taken out of the context of the citation (such as "Echocardiography has chemical part which affects Blood Flow Velocity"). The former case is of interest because this "common knowledge" was produced automatically. The latter case becomes clearer when additional relationships involving Echocardiography reveal that this procedure may include the use of certain chemicals and that these chemicals may affect Blood Flow. The types I, II, III and IV are described in the text.

### Type I: **X a Y**

The simplest type of relationship involves a direct semantic association between the two MeSH terms. For example, if a citation has "Electrocardiography" with the Subheading "Methods" and has "Myocardial Infarction" with the Subheading "Diagnosis", there is a rule which would propose the Type I relationship "Electrocardiography diagnoses Myocardial Infarction".

In some Type I relationships, the Subheadings involved can provide fairly specific information about the relationship (i.e., **a** can be quite explicit). For example, if a Disease is associated with a change in a (body) Chemical, the location (in the body) of the Chemical might be included in the relationship; thus, **a** might represent "is associated with blood change of" or "is associated with urine change of". This is possible due to the existence of Subheadings such as BLOOD and URINE. Thus, if a Disease (such as Aortic Valve Stenosis) appears in a citation with the Subheading BLOOD, and a Chemical (such as Neuroregulators) appears in the same citation with the Subheading ANALYSIS, the Type I relationship "Aortic Valve Stenosis is associated with blood change of Neuroregulators" would be proposed. The straightforward nature of Type I relationships makes them easy to generate. Of the 364 relationships, 199 (55%) were Type I relationships and, of these, 184 (92%) were valid; the remaining 15 (8%) were invalid.

### Type II: **X a *U* b Y**

A second class of relationships which can be extracted from MEDLINE citations involves two MeSH terms and some unnamed (unbound) third term. A common example is a Disease (Disease 1) which occurs during the treatment of a second Disease (Disease 2) with some Chemical. In the Type II syntax, **X** represents Disease 1, **Y** represents Disease 2, ***U*** represents "some chemical", **a** represents "is caused by" and **b** represents "treats". In the proposed relationship, the Chemical is not named. Instead, the two Diseases are indirectly related to each other through their direct relationships to the unnamed term.

Although there are three medical concepts involved with the Type II relationship, only two MeSH terms are required to satisfy the associated rule: those which correspond to **X** and **Y**. The unmentioned ***U*** may or may not appear in the citation, and relationships involving it specifically may be proposed by another rule (e. g., the third term serves as an **X** or **Y** in a Type I relationship). However, the presence of the third term in the citation is not required. For example, if "Aortic Valve Stenosis" appears in a citation with the Subheading "Physiopathology" and a chemical, such as "Technetium", appears in the citation with the Subheading "Diagnostic Use", there is a rule which will propose that the Type II relationship "Aortic Valve Stenosis is diagnosed by some procedure which has Technetium". The MeSH Procedure (presumably some nuclear scan) is not mentioned in the relationship. If the Procedure term appears in the citation with the appropriate Subheading, additional rules may be satisfied which propose Type I relationships about the procedure diagnosing Aortic Valve Stenosis or having the chemical part Technetium. However, the procedure may not be mentioned among the MeSH terms in the citation (perhaps no term for the procedure even exists in MeSH), making it impossible to propose any relationships other than the Type II relationship. Of the 364 relationships generated from the original search topics, 55 (15%) were Type II relationships. Of these 55 relationships, 53 (96%) were valid and 2 (4%) were invalid.

### Type III: **X** and **Y a *U***

In Type III relationships, the two identified terms share a common relationship with a third, unidentified term. (An equivalent syntactic description for Type III relationships is: **X a *U*** and **Y a *U***.) For example, two Diseases might be treated by the same Chemical. Once again, it is left to some other rules to detect the direct relationships (in this case, "treated by") between the Chemical and each Disease. It is entirely possible that the actual Chemical is not mentioned in the citation, making it impossible to find the direct relationships. Thus, if "Aortic Valve Stenosis" and "Heart Enlargement" appear in the same citation, each associated with the Subheading "Drug Therapy", the Type III relationship "Aortic Valve Stenosis and Heart Enlargement are treated by the same chemical" will be proposed. If the chemical is also included in the citation, and appears with the Subheadings "Pharmacology" or "Therapeutic Use", then other rules will propose the two Type I relationships concerning the Chemical and each of the Diseases. If the Chemical does not appear with a Subheading which satisfies a rule, or it does not appear at all, the Type I relationships obviously cannot be proposed, while rules proposing the Type III relationship will still be satisfied. Very few rules propose Type III relationships; this is reflected in the application of the rules to the Evaluation Set. Of the 364 relationships generated from the original search topics, 19 (5%) were Type III relationships. All 19 of these were valid.

### Type IV: **X a Y** and **Y a X**

In contrast to the previous three types of relationships, a Type IV relationship is one where no explicit or consistently reproducible association can be defined. Therefore, a Type IV relationship is not as informative because it indicates only nonspecific associations between pairs of MeSH terms. When the two terms are of the same class (both Diseases or both Procedures), a relatively loose semantic connection, such as "comparable" or "similar", exists between the terms. When the terms are of different classes, the association is merely "is related to". For example, if "Aortic Valve Stenosis" and "Myocardium" both appear in a citation with the Subheading "Metabolism", the Type IV relationship "Aortic Valve Stenosis is related to Myocardium" is proposed. Although this example might suggest a more specific relationship, such as "causes change in", the review of the training set revealed other Disease/Anatomic Site co-occurrences where the relationship was different (such as, "is caused by change in"). Thus, no consistently reproducible relationship exists, other than "is related to". Despite the minimal information provided by Type IV relationships, relative to the other three types, part of their strength lies in their numbers. There are many rules which produce them, making them easy to generate. Of the 364 relationships generated from the original search topics, 994 (26%) were Type IV relationships. All 94 of these were valid.

## 4. Discussion

Our method extracts facts, in the form of semantic relationships between MeSH concepts, from MEDLINE citations. While the method has proven fairly reliable for a narrow domain, the usefulness of the facts it generates is untested. However, it is possible to speculate as to their value, based on past and current research in the construction and use of medical knowledge bases. We include in this discussion several caveats for users of this approach and suggest topics for further research.

### *Rules Types and Automatic Knowledge Extraction*

There have been many efforts to develop tools for conducting automated knowledge acquisition. Approaches used for medical knowledge acquisition have included developing programs to work with domain experts [10, 11], to extract knowledge from clinical databases [12, 13], and to help physicians extract knowledge from medical literature [14]. The medical literature has been used more directly through specially-developed abstracts [4, 15]; our approach takes advantage of the readily available abstracts produced by the NLM.

```
[Myocardial Infarction/BLOOD AND /COMPLICATIONS]
                                OR
[Myocardial Infarction/CHEMICALLY INDUCED
    AND (/ADMINISTRATION & DOSAGE OR /ADVERSE EFFECTS OR /PHARMACOLOGY
        OR /POISONING OR /THERAPEUTIC USE)]
                                OR
[Myocardial Infarction/DIAGNOSIS AND (/COMPLICATIONS OR /THERAPY)]
                                OR
[Myocardial Infarction/DRUG THERAPY AND (/MORTALITY OR /THERAPY)]
                                OR
[Myocardial Infarction/ETIOLOGY
    AND (/ADVERSE EFFECTS OR /DIAGNOSIS OR /FAMILIAL & GENETIC OR /OCCURRENCE
        OR /PATHOLOGY OR /PSYCHOLOGY OR /RADIOGRAPHY OR /SECONDARY)]
                                OR
[Myocardial Infarction/FAMILIAL & GENETIC
    AND (/COMPLICATIONS OR /DIAGNOSIS OR /DRUG THERAPY OR /OCCURRENCE
        OR /PATHOLOGY)]
                                OR
[Myocardial Infarction/METABOLISM AND /DIAGNOSIS]
                                OR
[Myocardial Infarction/MORTALITY
    AND (/ADVERSE EFFECTS OR /COMPLICATIONS OR /ETIOLOGY OR /PATHOLOGY)]
                                OR
[Myocardial Infarction/OCCURRENCE
    AND (/ADVERSE EFFECTS OR /COMPLICATIONS OR /DIAGNOSIS OR /MORTALITY
        OR /PHYSIOPATHOLOGY)]
                                OR
[Myocardial Infarction/PATHOLOGY AND /COMPLICATIONS]
                                OR
[Myocardial Infarction/PHYSIOPATHOLOGY
    AND (/CONGENITAL OR /DIAGNOSIS OR /DRUG THERAPY OR/MORTALITY)]
                                OR
[Myocardial Infarction/PREVENTION & CONTROL
    AND (/COMPLICATIONS OR /ETIOLOGY OR /PHYSIOPATHOLOGY OR /THERAPY)]
                                OR
[Myocardial Infarction/REHABILITATION AND (/MORTALITY OR /THERAPY)]
                                OR
[Myocardial Infarction/SURGERY AND /COMPLICATIONS]
                                OR
[Myocardial Infarction/THERAPY AND /PHYSIOPATHOLOGY]
```

**Figure 2** Search Strategy for Causes of Myocardial Infarction. Upper-case terms are MeSH Subheadings; "OR" and "AND" are Boolean operators used to specify the search strategy. Note that searching for "Myocardial Infarction" produces a nonspecific set of citations. Searching for "Myocardial Infarction/ETIOLOGY" produces a more specific set; however, some citations may not list the etiologic agent among the indexed terms and this simple, straightforward strategy will fail to retrieve some relevant citations that will be retrieved by the search strategy shown in the Figure. While some parts of this strategy do not appear intuitive, they are derived from rules which were based on actual citations in which the disease and its cause appeared with these nonintuitive Subheadings. Note that by searching for "unattached" Subheadings, terms from various classes may be retrieved. For example, searching for "/THERAPEUTIC USE" will retrieve both Chemicals and Procedures. It is not possible to limit the search to only chemicals (i. e., the "explode" feature cannot be used with such a high-level term as "Chemicals"), so some procedures, which may not cause myocardial infarction, will be retrieved as well. However, the Rule Tables contain the information to determine which terms have the "is caused by" relationship to Myocardial Infarction.

It is, perhaps, easiest to imagine the usefulness of the Type I relationships in medical knowledge bases. For example, the DXplain knowledge base for medical diagnosis [2] could incorporate the fact that patients with Aortic Valve Stenosis might be predisposed to particular diseases and that patients with certain diseases might develop Aortic Valve Stenosis. Similarly, a knowledge base capable of suggesting useful diagnostic tests could make use of the list of 86 diseases diagnosed by Echocardiography.

The Type II relationships, being one logical step removed from Type I relationships, might be used by systems capable of some deeper level of reasoning. These relationships might be useful in expert systems which perform causal modeling (e. g., a chemical used in Echocardiography affects Blood Flow Velocity). They might also be useful in inferencing systems (for example, "If a disease affects the Aortic Valve then it might cause Aortic Valve Stenosis").

Similarly, Type III relationships might be used in interesting ways by inferencing systems which suggest alternative diagnostic or therapeutic strategies. For example, the evaluation set provides the relationship "Aortic Valve Stenosis and Mitral Valve Stenosis are treated by the same procedure". An inferencing mechanism would be capable of proposing some potentially interesting treatment strategies, such as "If a procedure is known to treat Aortic Valve Stenosis it might also treat Mitral Valve Stenosis". Use of this rule in combination with another relationship proposed from the evaluation set, such as "Aortic Valve Stenosis is treated by Heart Surgery" would allow the inferencing mechanism to suggest, "Mitral Valve Stenosis might be treated by Heart Surgery".

The Type IV relationships provide information about terms that are simply "related" to each other. Typically, many MeSH terms appear in a MEDLINE citation; relationships exist for some pairs of terms, while others are merely coincident. Rules which propose Type IV relationships permit the differentiation between related terms and coincident terms. Classification as a Type IV is done when no more specific relationship can be identified. There is potential value in these Type IV relationships, despite their generality. An expert system which attempts to classify terminology might make use of information provided by the Type IV relationships. One such system categorizes medical terms based on their occurrence in journal abstracts [3], while another technique, called sublanguage analysis, has been used to classify terms based on natural language processing [16]. Additional information about the classification of terms can be obtained from Type IV relationships gleaned from MEDLINE citations. These relationships, in turn, can form the basis for a classification scheme. For example, if a survey of citations reveals that "Endocarditis is related to Myocarditis", "Myocarditis is related to Aortitis", and "Aortitis is related to Arteritis", we have detected a new class of terms (which we might call "inflammatory cardiovascular diseases").

### *Applications for Knowledge Extracted from MEDLINE*

One potential area where knowledge extraction can be applied is in literature retrieval. Powsner, et al. have previously studied the semantic relationships in medical literature and examined MEDLINE citations for patterns in the Headings and Subheadings which might serve to improve literature retrieval [7]. Using a different approach, the method described here also serves to produce patterns which may be useful for formulating search strategies. Such patterns could be used by systems which guide users in literature searches, such as Grateful Med® [17] or MicroMeSH® [18]. If there are multiple rules which propose a relationship of interest to a user, those rules could be converted into search requests. For example, if a rule which proposes causes for a disease is of the form:

If Disease/Subheading 1 AND
MeSH Term/Subheading 2
THEN MeSH Term causes Disease,

then a possible search strategy to retrieve articles about the causes of the disease would be:

Disease X/Subheading 1 AND
Subheading 2.

Since there are many rules capable of proposing this relationship, the search can be augmented by combining all the appropriate rules to produce a search such as:

(Disease X/Subheading 1 AND
Subheading 2) OR
(Disease X/Subheading 3 AND
Subheading 4) OR
(Disease X/Subheading 5 AND
Subheading 6) OR..

Consider a specific example. If we were interested in retrieving all the causes of Myocardial Infarction, we might use the search strategy "Myocardial Infarction/ETIOLOGY"; however, the Rule Tables can suggest the much more complex search strategy shown in Figure 2. Searching one particular MEDLINE database for "Myocardial Infarction" retrieved 14,541 citations (many of which did not deal with the cause of myocardial infarction) while a search of the same database for "Myocardial Infarction/ETIOLOGY" retrieved 1,320 citations. The strategy shown in Figure 2 retrieved 3,852 citations (including 3,013 not found by the "Myocardial Infarction/ETIOLOGY" search), all of which listed the etiologic agent among the indexed terms. For example, one citation was retrieved that contained "Myocardial Infarction/BLOOD" and "Pyridoxine Deficiency/COMPLICATIONS". None of the citations indexed with "Myocardial Infarction/ETIOLOGY" reference "Pyridoxine Deficiency", so the intuitive search would have failed to find an article [19] that provides knowledge about a potential factor in the cause of myocardial infarction. More research will be needed before the utility of such rule table-derived search strategies can be assessed.

Another potential use for automatically generated semantic relationships (and, in fact, the one that inspired this work) is in the construction of semantic networks. A semantic net is an expressive means for representing knowledge [20]. If a semantic net were to be constructed with MeSH terms as nodes, then semantic relationships between terms in MEDLINE citations could be represented by slots and links. In addition to the semantic links described here, studies that are part of the NLM's Unified Medical Language System (UMLS) effort have examined the semantic links which are meaningful between classes of medical concepts, which include MeSH terms [21–23]. For example, if the semantic network contains a node for "Aortic Valve Stenosis", one of its slots might be "Affected Body Parts" which contains the MeSH term "Aortic Valve". This reference is, in effect, a semantic relationship from the first MeSH term to the second. Similarly, the network node for "Aortic Valve" might include a slot called "Diseases" which contains a reference (back) to "Aortic Valve Stenosis". Although this is a seemingly trivial example, it would seem that a valuable semantic network could be constructed from such relatively simple connections. Manual compilation of all of the useful simple connections could be enormously time-consuming. An approach such as the one described could identify many of the useful connections automatically.

Automatically generated semantic links might also help alleviate another tedious task: construction of hypertext links [23]. We have explored the potential usefulness of these semantic relationships by incorporating them into a medical hypertext environment so that readers can traverse semantic links among terms in the text and, owing to the means by which the relationships are created, can access relevant medical literature citations [24].

### *Caveats for Building Knowledge Bases from Hypotheses*

In all these discussions, it is important to recall that the relationships generated by this approach are merely propositions. In all cases, the validity of each must be verified by a medical expert. This should be a fairly simple process, since the expert has prior knowledge which can corroborate many of the propositions which are generated (e. g., "aortic valve stenosis affects the aortic valve"). The automated extraction of relationships from MEDLINE can produce large quantities of information much more easily than a knowledge engineer extracting it from an expert [25]. Furthermore, for those proposed relationships which the expert disputes or cannot corroborate, the process provides support for each of its propositions with literature citations, since literature citations caused the relationship to be proposed. The expert may then review this literature to see if the relationship is borne out by the citation. The proposed relationships for Aortic Valve Stenosis and Echocardiography were evaluated through such a review. During that review, four situations were identified in which the proposition could be inappropriate or misleading.

The first example where knowledge, derived by the proposed automatic extraction techniques may be misleading, is due to the fact that the rules were constructed by examining a subset of the entire MEDLINE database (less than 0.1% of the citations in the database). Thus, there is no guarantee that the rules will be valid if applied to all citations. The process by which the Rule Tables were constructed was an iterative one which eventually yielded a rule base that was "consistent across the training set". In fact, during this process, 12 rules which were initially believed to be valid were later discovered to be inconsistent when applied to the new citations which were contained in the evaluation set. As the method is applied to new domains (such as non-cardiac diseases and procedures), additional rules can be expected to fall by the wayside (although new rules may also be discovered).

A second cause of misleading propositions is that the propositions extracted by the rules are, of course, only as good as the citations from which they are drawn. Indexing is a manual process, so the potential for error always exists and this would obviously affect the results. Of the 3,056 citations reviewed, only one indexing term was noted to be used in a somewhat inappropriate, albeit consistent, manner: several articles dealing with a diagnostic technique known in Europe as a "nuclear stethoscope" (a radionuclide imaging procedure which neither detects sound nor makes use of a traditional stethoscope) were indexed with the term "Heart Auscultation". This yielded some peculiar results (e. g., "Heart Auscultation has chemical Technetium"). It can be expected that such deviations in MeSH term usage affect the results of the knowledge extraction.

A third potential source of inappropriate results could occur even when the rules and citations are correct. Journal articles which are later refuted remain in the MEDLINE database. (Even those articles which have been officially retracted by their authors for scientific misconduct remain in the corpus of the electronic literature, although they are flagged in MEDLINE as having been retracted.) If the knowledge reported in the medical literature is incorrect, especially in areas of controversy, the erroneous information still passes into the citations and will be extracted. A single example of this occurrence was encountered: "Internal Mammary Artery Implantation" was proposed as a treatment for myocardial infarction. At the time that the cited literature was published, this was believed to be true; since then, it has been learned that this procedure is ineffective. This problem can be alleviated in part by ignoring articles flagged as retracted and by maintaining surveillance for subsequent retractions. Since individual relationships in the knowledge base are traceable to original citations, identification of a retracted article can lead to retraction of the relationships derived from the citation.

Finally, even when the literature is correct, the citation is composed accurately and the rule base is sound, that small portion of the medical literature which is trivial or humorous may produce trivial or humorous results. A favorite example of this was the proposition "Syncope is diagnosed by Tooth Extraction". It derives from a letter written by a dentist who successfully diagnosed his patient's syncope by performing electrocardiograms during dental procedures. The citation included "Syncope/Diagnosis" and "Tooth Extraction/Methods", which triggered a rule, generating the peculiar proposed relationship. In fact, the abstracter used the Headings and Subheadings appropriately and the rule base interpreted the citation correctly: tooth extraction was being used as a form of provocative test. The parts of the process were well correct; nevertheless, this is not a particularly useful approach to the diagnosis of syncope and is not likely to be a welcome addition to expert systems lacking a sense of humor.

### *Other Areas of Research*

Only 10% of possible Subheading permutations, 0.3% of MeSH terms, 0.05% of the MEDLINE database, and one general pattern type (Term/Subheading AND Term/Subheading) have been examined. The initial rule set was based on the domain of two cardiovascular diseases and two cardiovascular procedures (syncope, myocardial infarction, heart auscultation, and angiocardiography), while evaluation was one in the domain of a different cardiovascular disease and procedure (aortic valve stenosis and echocardiography). Given the differences between these two narrow domains, our results suggest that this method should be appropriate for the more general domain of cardiovascular diseases and procedures. Only further evaluation can determine if the method is indeed generalizable or transferable to other domains. The almost-minimal rule form of "If A AND B THEN C" presents only one type of pattern present in MEDLINE citations. Rules with three or more conditions might also prove useful. For example, examining the MeSH "check tags" (such as HUMAN) which appear in the citations, might improve specificity of some rules. More complex rule structures might also be of value, since they might detect more complex relationships (e. g., "Chemicals A, B and C can be compared in the treatment of Disease D").

Another area which could be explored is the "certainty factor" of the results. One way to strengthen the certainty of a proposed relationship, for example, might be to tally the number of citations which support the proposition. If this approach were taken, rules which were heretofore demand "inconsistent" (because of one incorrect result) and discarded might be salvaged with the understanding that their specificity was somewhat less than 100%. This diminution in specificity would be offset by a corresponding increase in sensitivity which might be valuable with regard to knowledge extraction.

Maintenance issues also bear some exploration. Knowledge bases constructed by our method can be maintained by ongoing processing of newly-published literature with review of suggested additions. The rule base itself can also be maintained through a pair of complementary procedures. First, as new literature is published and indexed in MEDLINE, new patterns of MeSH terms-Subheading co-occurrences will be encountered (corresponding to the 5,502 empty cells referred to in Table 1). These patterns can suggest additions to the Rule Table using the same methods that were applied to the original training set. Second, rules can be removed from the Rule Table by noting the proposal of incorrect relationships based on processing of new literature citations.

As this approach is extended to other areas of the medical literature, new Subheadings and new classes of MeSH terms might enable the construction of new semantic relationships. For example, if a training set were to be created based on a literature search for infectious diseases, it is likely that new rules would be created which make use of heretofore unused Subheadings, such as MICROBIOLOGY, PARASITOLOGY, PATHOGENICITY and TRANSMISSION.

It is also likely that the relationships proposed would involve a new class of terms: Organisms.

## 5. Conclusion

Our work demonstrates that it is possible to extract medical knowledge from the world literature in an automated way by taking advantage of implied knowledge incorporated in MEDLINE citations by the NLM indexers. This knowledge takes the form of semantic relationships among medical concepts. The relations are grouped into four syntactic types, which are represented with a simple grammar.

The "knowledge" which can be generated automatically in this manner may be misleading because of the possibility of four different ways in which error can creep into the results. However, we believe that the system can act as an automated research assistant to serve those who seek to collect medical knowledge for inclusion in knowledge bases. This method has an advantage in that the proposed knowledge comes with ready-made supporting references to the medical literature. This method will also be useful for those who maintain knowledge bases, by continually applying the rule base to searches restricted to the most recent literature.

We believe that the knowledge that the system generates will be particularly useful to those constructing semantic networks, such as the one developed for our own medical hypertext environment [24] or for the UMLS [22]. The automated extraction of knowledge, directly from the medical literature, is a desirable but elusive goal. One attractive step is to take advantage of the structure and controlled vocabulary in the MEDLINE database.

**Acknowledgements**

This work was supported in part by NLM Training Grant T15-LM07037-4, UMLS Contract N01-LM-8-3513, and an equipment grant from the Hewlett-Packard Corporation. The authors wish to acknowledge Dr. Henry J. Lowe, M. D. for the data structures used to manage the MeSH terms and Ms. Laurie J. Hassan for invaluable programming assistance. The authors thank Ms. Leslie Juceam and Dr. George Hripcsak for editorial advice.

REFERENCES

1. Miller RA, Pople HE, Myers JD. INTERNIST-I: An experimental computer-based diagnostic consultant for general internal medicine. N Engl J Med 1982; 307: 468–76.
2. Barnett GO, Cimino JJ, Hupp JA, Hoffer EP. DXplain – An evolving diagnostic decision-support system. J Am Med Assoc 1987; 258: 67–74.
3. Cole WG, Michael PA, Stewart JG, Blois MS. Automatic classification of medical text: The influence of publication form. In: Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care. Greenes RA (ed). New York: IEEE Computer Society Press, 1988: 196–200.
4. Rennels GD, Shortliffe EH, Stockdale FE, Miller PL. A computational model of reasoning from the medical literature. Comput Meth Prog Biomed 1987; 24: 139–49.
5. Bachrach CA, Charen T. Selection of MEDLINE contents, the development of its thesaurus, and the indexing process. Med Inform 1978; 3: 237–54.
6. National Library of Medicine, Library Operations. Medical Subject Headings. Bethesda, Md: National Library of Medicine, 1992.
7. Powsner SM, Riely CA, Barwick KW, Morrow JS, Miller PL. Automated bibliographic retrieval based on current topics in hepatology: Hepatopix. Comput Biomed Res 1989; 22: 552–64.
8. Cimino JJ, Mallon LJ, Barnett GO. Automated extraction of medical knowledge from MEDLINE citations. In: Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care. Greenes RA (ed). New York: IEEE Computer Society Press, 1988: 180–4.
9. Harris Z. Language and Information. New York: Columbia University Press, 1988.
10. Walton JD, Musen MA, Combs DM, Lane CD, Shortliffe EH, Fagan LM. Graphical access to medical expert systems: III. Design of a knowledge acquisition environment. Meth Inform Med 1987; 26: 78–88.
11. Shahsavar N, Gill H, Wigertz O, Frostell C, Matell G, Ludwigs U. KAVE: a tool for knowledge acquisition to support artificial ventilation. Comput Meth Prog Biomed 1991; 34: 115–23.
12. Warner HR, Haug P, Bouhaddou O, Lincoln M, Warner H Jr, Sorenson D, Williamson JW, Chinli F. ILIAD as a expert consultant to teach differential diagnosis.

In: Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care. Greenes RA (ed). New York: IEEE Computer Society Press, 1988: 371–6.

13. Chowdhury S, Bodemar G, Haug P, Babic A, Wigertz O. Methods for knowledge extraction from a clinical database on liver diseases. Comput Biomed Res 1991; 24: 530–48.
14. Giuse DA, Giuse NB, Miller RA. Towards computer assisted maintenance of medical knowledge bases. Artif Intell Med 1990; 2: 21–33.
15. Giuse DA, Giuse NB, Miller RA. Heuristic determination of quantitative data for knowledge acquisition in medicine. Comput Biomed Res 1991; 24: 261–72.
16. Johnson SB, Gottfried M. Sublanguage analysis as a basic for a controlled medical vocabulary. In: Proceeding of the Thirteenth Annual Symposium on Computer Applications in Medical Care. Kingsland LC (ed). New York: IEEE Computer Society Press, 1989: 519–23.
17. Lindberg DA, Schoolman HM. The National Library of Medicine and medical informatics. West J Med 1986; 145: 786–90.
18. Lowe HJ, Barnett GO, Scott J, Mallon L, Ryan-Blewett D. Remote Access MicroMeSH: Evaluation of a microcomputer system for searching the MEDLINE database. In: Proceedings of the Thirteenth Annual Symposium on Computer Applications in Medical Care. Kingsland LC (ed). New York: IEEE Computer Society Press, 1989: 445–7.
19. Kok FJ, Schrijver J, Hofman A, et al. Low vitamin B6 status in patients with acute myocardial infarction. Am J Cardiol 1989; 63: 513–6.
20. Barr CE, Komorowski HJ, Pattison-Gordon E, Greenes RA. Conceptual modeling for the Unified Medical Language System. In: Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care. Greenes RA (ed). New York: IEEE Computer Society Press, 1988: 148–51.
21. Humphreys BL, Lindberg DAB. Building the Unified Medical Language System. In: Proceedings of the Thirteenth Annual Symposium on Computer Applications in Medical Care. Kingsland LC (ed). New York: IEEE Computer Society Press, 1989: 475–80.
22. McCray A. The UMLS Semantic Network. In: Proceedings of the Thirteenth Annual Symposium on Computer Applications in Medical Care. Kingsland LC (ed). New York: IEEE Computer Society Press, 1989: 503–7.
23. Greenes RA, Tarabar DB, Krauss M, et al. Knowledge management as a decision support method: A diagnostic workup strategy application. Comput Biomed Res 1989; 22: 113–35.
24. Cimino JJ, Elkin PL, Barnett GO. As we may think: the Concept Space and medical hypertext. Comput Biomed Res 1992; 25: 238–63.
25. Musen MA, van der Lei J. Knowledge engineering for clinical consultation programs: Modeling the application area. Meth Inform Med 1989; 28: 28–35.

Address of the authors:
James J. Cimino, M. D.,
Center for Medical Informatics,
Atchley Pavilion – Room 1310,
Columbia-Presbyterian Medical Center,
New York, New York 10032,
USA