Int J Digit Libr (1997) 1: 288–296

INTERNATIONAL JOURNAL ON
Digital Libraries
© Springer-Verlag 1997

Medical information retrieval

Development and evaluation of a context-based document representation for searching the medical literature

Gretchen P. Purcell[1], Glenn D. Rennels[2], Edward H. Shortliffe[2]

[1] Department of Surgery, Duke Hospital North Rm 3423, Duke University Medical Center, Durham, NC 27710, USA
[2] Section on Medical Informatics, Stanford University School of Medicine, Stanford, CA 94305, USA

Received: 12 March 1997 / Revised: 26 June 1997 / Accepted: 5 August 1997

Abstract. Conventional full-text systems represent documents as sets of index terms, and queries to these systems often retrieve irrelevant material when search terms occur in inappropriate contexts. We have developed document representations that capture the semantic contexts in which text words occur. Many bodies of literature contain stereotypic categories of information. For example, articles describing medical research consistently discuss interventions and outcomes. These semantic themes provide context for terms in the text, and thus, can facilitate precise full-text searches. We have used a contextual model of clinical research articles, case reports, and review articles as the basis for a document representation in a full-text retrieval system. In this paper, we describe the creation of context models for medical publications and the evaluation of these models using interindexer consistency. We demonstrate that such models are easily understood and employed by readers of the literature (and thus, the searchers). Accordingly, these models may constitute a powerful representation for information retrieval. We discuss the suitability of this technique for other domains.

Keywords: Medical publications – Context models – Full-text retrieval – Interindex consistency

1 Introduction

Publishers of scientific and technical journals, newspapers, and magazines now provide electronic access to their materials [11, 17]. Networks that connect libraries and universities expand the literature available to students, researchers, and the community [7, 24]. Traditional search tools have not evolved to accommodate this increasing volume of textual information.

In conventional searching systems, descriptors known as index terms represent the content of documents. Information-retrieval systems match query terms with index terms to identify relevant information [31]. Imperfect methods for the selection of index terms produce both inaccurate and inadequate representation of documents, and thus, compromise retrieval performance.

Many information-retrieval systems draw index terms from large, controlled vocabularies. For example, the National Library of Medicine (NLM) selects index terms for the biomedical literature database, MEDLINE, from the Medical Subject Headings (MeSH) vocabulary, which contains over 18 000 concepts. This terminology undergoes hundreds of changes each year to reflect evolving scientific knowledge [26, 27]. Clinicians struggle to use such enormous searching vocabularies effectively [23]. Even professional indexers cannot consistently apply the MeSH vocabulary in assigning document descriptors [6, 18], and information scientists have demonstrated that inconsistent indexing undermines the performance of retrieval systems [20].

The NLM has developed concept hierarchies, such as the Metathesaurus of the Unified Medical Language System, to aid users of information systems in managing controlled vocabularies [21, 32]. However, retrieval systems that depend on such hierarchies cannot always perform optimally because evolving topics may not be incorporated in a timely manner [9, 10]. Users are likely to seek information about emerging disciplines, but they may be unable to retrieve relevant documents if the indexing vocabulary does not represent new concepts. Finally, limitations on the number of index terms can also render a document representation inadequate. The 10–12 index terms that represent a MEDLINE document often do not capture all topics in an article that might be of interest to a searcher.

Full-text retrieval systems select index terms from the text of a document, and thus provide a more robust representation of concepts [33]. However, these systems still suffer from the shortcomings of term-based document representations. Algorithms for automatic indexing choose index terms based on measures such as word frequency or the ability of a term to distinguish one document from others [31]. Statistical measures cannot reflect the importance or meaning of a term in a document. Thus, full-text retrieval systems often overwhelm users with irrelevant material.

We have observed that full-text retrieval systems often return irrelevant documents that contain search terms in an inappropriate context. For example, a physician who seeks information about the treatment of high blood pressure may retrieve articles that discuss the term "hypertension" as a risk factor for stroke or as a complication of pregnancy. Literature searches erroneously identify these documents because term-based representations cannot capture such distinctions in usage. The meaning of an index term depends on the context in which it is used.

In this paper, we present a document representation that includes words from a document and the contexts in which those terms appear. The combination of a robust full-text indexing scheme and a contextual structure that narrows the possible interpretations of index terms provides the foundation for precise retrieval systems.

2 Context-based document representations

A **context** is a characterization of a text that provides the basis for understanding potentially ambiguous terms or phrases. For example, "systolic blood pressure greater than 180" has several possible interpretations. In the discussion of eligibility for a clinical trial, this phrase might represent a criterion for enrollment in a study of anti-hypertensive drugs. The same phrase, in the context of exclusion criteria, could specify a reason for withdrawing patients from a study of thrombolytic therapy. Alternatively, the phrase might indicate a condition that predisposes patients to coronary artery disease in a discussion of risk factors. In these three contexts, the interpretations of the same textual phrase differ significantly. Similarly, such contextual information can elucidate the meaning of individual terms in a text.

Articles from biomedical journals consist of three general types: 1) papers that report data, including original research articles and case presentations, 2) didactic papers, such as review articles and teaching cases, and 3) papers that comment or speculate, including editorials and letters [30]. Within each type of publication, authors present characteristic classes of information. For example, research articles contain methods and results, and teaching cases describe clinical scenarios and possible diagnoses. Instructions for writing abstracts and papers outline the elements of scientific articles that have defined biomedical publications for decades [1, 2, 13, 14, 25].

In our document representation, a **context model** enumerates the recurring semantic themes or contexts in a publication. Figure 1 presents the context model for clinical research articles. We associate each sentence from a document with one or more contexts from the model. For example, both the *Objective* and the *Study type* contexts would characterize the following sentence: "We conducted a randomized controlled trial to evaluate the effects of exercise on blood pressure." Terms from a text are indexed according to the contexts in which they occur. Thus, a clinician might seek articles with "randomized" in the *Study type* context. We call this technique **context-based searching**.

Clinical research article
- Title
- Authors
- Background
- Objective
- *Methods*
  - Study Type
  - Study Setting
  - *Study-population methodology*
    - Eligibility / Selection
    - Exclusion / Withdrawal
    - Study-Group-Assignment Procedures
  - *Experimental parameters*
    - Risk/Prognostic Factor Assessment
    - Intervention Evaluated
    - Concomitant Interventions
    - Diagnostic Test Evaluated
    - Gold Standard
    - Outcome Measures / Endpoints
  - Consent / Ethics Procedures
  - Statistical Methods
- *Results*
  - Experimental Findings
  - Adverse Effects
- Conclusions
- Limitations / Biases
- Future Work
- Acknowledgments / Collaborators
- References

Fig. 1. The context model for clinical research articles. This figure outlines the contexts that characterize clinical research articles. Compound contexts are underlined and capitalized. Component contexts follow their parent contexts with one level of indentation

Contexts models are organized hierarchically as shown in Fig. 1. **Compound contexts**, such as *Study-population methodology* represent classes of contexts that share common features. The top-level context, *Clinical research article*, is a publication type.

Context markup is the process of assigning contexts to sentences in a document. In our implementation, this task involves inserting context-specific tags around appropriate portions of a text. Once contexts have been specified in collection, a full-text indexing algorithm employs this markup to generate context-based indices.

Context markup does not affect the appearance of a clinical publication. Formatting conventions, such as the structured abstract, dictate the layout of a document [1, 13, 25]. Such constructs are not ideal for searching, however, because they do not faithfully reflect the content of the text [5]. Even common section headings, such as Methods and Results, do not accurately characterize the associated content. Authors of research articles routinely report experimental findings under the Methods heading, and discuss study methodology in the Results section. Such writing styles may be necessary to explain complex experiments. Structured formats can interfere with the author's ability to compose comprehensible prose [8].

The contextual structure accurately characterizes the text without influencing the presentation. Sentences that are associated with a particular context may occur in different paragraphs or on different pages of a document. For example, clinical research articles present BACKGROUND information in the beginning of a paper to motivate the research, and with the results to compare with previous work. Context markup captures this semantic information in a manner that is independent of the document format or organization.

We have developed context models for clinical research articles, didactic case reports, and reviews. Figures 1, 2 and 3 illustrate these models, respectively. Research articles present the methods, results, and implications of a clinical study. Case reports are articles that describe a clinical scenario, discuss the possible diagnoses and treatment options, and present the most appropriate management. Reviews summarize current knowledge in clinical medicine or research. Context models represent the recurring themes in each of these medical publications.

We have implemented an information-retrieval system that employs context models in searching the medical literature. In related work [28], we have demonstrated that the context model for research articles can significantly improve the precision of full-text searching at fixed levels of recall. We have also constructed tools that employ the contextual structure to extract information from clinical research articles for specific clinical tasks [29].

These benefits in searching and displaying the medical literature require an explicit contextual structure created by context markup. The markup process is an indexing task. Manual indexing can require significant expertise, and it can be time consuming and inconsistent. To be useful, an indexing scheme should be easily applied by indexers and understood by the expected searchers [31]. In the following section, we describe an evaluation of the context models as an indexing scheme.

Case report
- Title
- Authors
- Discussants
- *Case presentation*
  - Case Findings
  - Case Management
- *Case discussion*
  - *Discussion of diff•rential diagnosis*
    - Ddx - Diseases in Differential Diagnosis
    - Ddx - Diagnostic Test Implications
    - Ddx - Management Implications
    - Ddx - Prognostic Implications
  - *Discussion of definitive/presumed diagnosis*
    - Dx - Definitive/Presumed Diagnosis
    - Dx - Diagnostic Test Implications
    - Dx - Management Implications
    - Dx - Prognostic Implications
- Acknowledgments
- References

Fig. 2. The context model for case reports. The two main components of a case report are the presentation of a specific clinical problem and a didactic discussion of the case. The case discussion consists of the differential diagnosis for the case and a discussion of the actual diagnosis and management plan

Review article
- Title
- Authors
- Objective
- Background
- *Reviewed topics*
  - Reviewed Population or Diseases
  - Reviewed Interventions
  - Reviewed Risk Factors / Exposures
  - Reviewed Tests
  - Reviewed Outcomes
  - Reviewed Adverse Effects / Complications
- *Relevant topics*
  - Relevant Population or Diseases
  - Relevant Interventions
  - Relevant Risk Factors / Exposures
  - Relevant Tests
  - Relevant Outcomes
  - Relevant Adverse Effects/Complications
- *Review methodology*
  - Data-Source Selection Procedures
  - Data-Source Exclusion Procedures
  - Data-Extraction / Statistical Methods
- *Review results*
  - Review Data
  - Sources of Variation
  - Biases / Limitations
- Conclusions
- Future Work
- Acknowledgments
- References

Fig. 3. The context model for reviews. This model characterizes articles that provide an overview of medical knowledge or research. The *Review methodology* determines the literature to be reviewed, and the *Review results* summarizes the findings. The focus of the review and topics of secondary interest are represented by the *Reviewed topics* and *Relevant topics* contexts, respectively

3 Methods for the evaluation of context models

Information scientists evaluate indexing schemes and the indexing process with experiments that measure the ability of different people to reproduce the indexing for a set of documents [6, 18]. Indexing consistency correlates with retrieval-system performance [20]. Thus, studies of interindexer consistency measure the contribution of an indexing scheme to the performance of a searching tool. To evaluate our document representation, we conducted studies of interindexer consistency for each context model. Because the indexing process is context markup, we refer to our experiments as studies of **intermarker consistency**.

An effective indexing scheme is easily understood by the searchers of a retrieval system [31]. Thus, we recruited the expected users of context-based searching, instead of professional indexers, as experimental subjects. Physicians, nurses, residents, interns, and medical students were eligible, and study groups were assembled from available volunteers. We excluded anyone who had assisted in the development of the context models. Each subject participated in the evaluation of only one context model.

We provided approximately two hours of training prior to each context-markup experiment. Subjects were given a description of the context model with a definition for each context. An instructor explained the context model and the context markup process using a marked article as an example. Each subject practiced context markup on three to five journal articles during the training session. The instructor distributed solutions and discussed areas of confusion. We permitted the subjects to refer to the training materials during the intermarker consistency studies.

We selected articles for both the training sessions and the experiments from three widely read and respected American journals for internal medicine: *Annals of Internal Medicine*, *Journal of the American Medical Association*, and the *New England Journal of Medicine*. Topics in internal medicine are of general interest among students, residents, and clinicians, and we chose this domain to aid with the recruitment of experimental subjects. Because only the *New England Journal of Medicine* contained teaching case reports, we also included articles from the *British Medical Journal* in the evaluation of this context model. In the training sessions, we employed articles that illustrated features of the context models. For each study of intermarker consistency, we systematically chose five articles from the most recent issues of the four journals that were available in our laboratory. We excluded papers that had been analyzed during context-model development to avoid bias in the training. We attempted to represent each journal equally in the experiments. Because of the time-consuming nature of the studies, we excluded unusually long manuscripts.

Each subject marked paper copies of three to five test articles with contexts from one of the models. We provided a numbered list of contexts and the following instructions for the experiments:

1. Please mark all sections of the document with one or more contexts except figures, tables, graphics, and section headings. *Do not forget to mark the title, authors, references, and discussants.*
2. Mark each sentence or block of text with a *number* and a *name* or abbreviation for the appropriate context.
3. Do not divide sentences when marking the text with contexts. If more than one context applies to a sentence, indicate both contexts.

We directed the subjects to select the most specific context that was applicable to a sentence, and thus, did not permit markup using compound contexts. Subjects could mark the test articles at any time after the training session, but we encouraged them to complete the experiment within one week of the instruction.

We measured intermarker consistency with the kappa coefficient of agreement for nominal scales [3]. The kappa statistic determines the reproducibility of assignments to unordered or nominal categories. The equation for kappa, shown in Fig. 4, calculates the proportion of agreement which exceeds the agreement expected by chance among a set of judges. In our experiment, the judges were experimental subjects who assigned sentences to contexts.

To assess the strengths and weaknesses of our contexts models, we computed a kappa value for each context. For a given sentence and a pair of experimental subjects, agreement was achieved when either: 1) both subjects assigned a context to the sentence, or 2) both subjects did not assign the context to the sentence. Mean agreement was calculated over all possible pairs of context markers across all sentences in the test documents. We estimated the agreement due to chance using the method described by Fleiss [4].

$$K = \frac{p_{agree} - p_{chance}}{1 - p_{chance}}$$

p_{agree} = Proportion of trials in which judges agree

p_{chance} = Proportion of trials in which agreement would be expected due to chance

Fig. 4. The kappa coefficient of agreement. This equation measures the fraction of beyond-chance agreement achieved by a set of judges who have assigned objects to unordered categories

We calculated an **overall kappa** to measure the agreement of the markup process as a whole. This summary statistic is the average of individual kappa values weighted by the frequency of their use during context markup. The **context frequency** is the total number of sentences marked with a context divided by the total number of context assignments. Figure 5 shows the calculation of an overall kappa.

Benchmarks for the interpretation of the kappa statistic are defined in the literature (Table 1; [19]). The subjects of the intermarker experiments received only two hours of training, and thus, we did not expect to observe perfect agreement for all contexts. We sought to achieve moderate agreement (kappa values greater that 0.40) overall and for one half of the individual contexts in each model. We believe that this level of agreement across novice context markers provides evidence for an under standable and reproducible indexing scheme. We

$$K_{overall} = \frac{\Sigma \; K_i \; f_i}{\Sigma \; f_i}$$

K_i = Kappa coefficient for context i

f_i = Frequency of context i

Fig. 5. The overall kappa. This equation determines the average beyond-chance consistency with which contexts were assigned to a set of documents. The overall kappa weights the kappa values for individual contexts by their frequency in the document. The sum of weighted kappa values is divided by the sum of context frequencies

Table 1. Interpretation of the kappa coefficient

Kappa statistic	Interpretation
< 0.00	Poor
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect

employed pilot studies of intermarker consistency in developing the context models. During the pilot studies, each subject marked only three articles, and we permitted markup with compound contexts because the models were often incomplete. These preliminary experiments identified missing, extraneous, and confusing contexts. Based on the results, we modified the context models and repeated the studies of intermarker consistency. In testing the revised models, we prohibited the use of compound contexts.

In this paper, we present the second and final studies of intermarker consistency for the revised models of clinical research articles and case reports. The model for review articles required substantial modification, and thus, we present only the results of the pilot study for this publication.

4 Results

In studies of intermarker consistency for clinical research articles, teaching case reports, and review articles, we observed moderate or better levels of intermarker consistency. Overall kappa values were 77 percent, 74 percent, and 52 percent, respectively. Table 2 summarizes these results.

If we remove the contributions of easily identified contexts (i.e., *Title*, *Authors*, *Discussants*, and *References*), we still demonstrate substantial intermarker consistency for both clinical research articles and case reports (overall kappa values of 71 percent and 68 percent, respectively). With these contexts omitted from the analysis of review articles, the intermarker consistency is only fair (overall kappa value of 30 percent). The *References* context applied to over one third of the reviews in our experiments, and thus, the agreement for this context contributed substantially to the kappa value. Overall agreement in the body of the text was fair to poor.

For clinical research articles and case reports, our observed levels of indexing consistency are generally better than the percentage agreement observed in studies of interindexer consistency for the MEDLINE database [6, 18]. Although MeSH indexing differs from context markup, we can informally compare our respective findings. The assignment of MeSH terms to articles from the MEDLINE database represents the state-of-the-art in human indexing; the professional indexers who perform this task train for at least 1 year [12]. In an interindexer-consistency study of 760 MEDLINE journal articles that were coincidentally indexed twice, the mean agreement ranged from 33.8 to 74.7 percent [6]. These values do not account for agreement due to chance. Because the MeSH vocabulary contains over 18 000 terms, chance agreement is unlikely. However, MEDLINE indexers achieved the best interindexer consistency (74.7 percent) for checktags, a small set of descriptors that are checked for applicability to all articles. Chance agreement probably accounted for a significant part of this observed consistency. Agreement for the index terms that reflect the most important themes in an article (central-concept main headings) was 61.1 percent; the value for all MeSH terms assigned to an article was only 48.2 percent. Thus, the high levels of intermarker consistency that our novice context markers achieved are comparable or better than the agreement of professional MEDLINE indexers.

Table 2. Overall kappa values

Context model	Overall kappa	Interpretation
Clinical research articles	0.77	Substantial
Case reports	0.74	Substantial
Review articles	0.52	Moderate

Three clinical medical students and two residents, one each in pediatrics and radiation oncology, participated in the study of intermarker consistency for clinical research articles. The test collection comprised two trials of therapeutic interventions, two studies of diagnostic tests, and one article about risk factors. Table 3 presents the results of this experiment. For each context, Table 3 reports the kappa value, the interpretation of the degree of agreement according to the Landis benchmarks, and the frequency of the context in the test collection.

The context markers achieved moderate or closer agreement in marking 80 percent of the contexts in the model. For several contexts, low frequencies in the test collection might account for disagreement among context markers. For example, only slight agreement was observed for the contexts *Study-group Assignment* and *Concomitant interventions* that rarely occurred. The frequency of the inconsistently marked *Prognostic/risk factor assessment* context was greater, but sentences that were associated with this context occurred in only one of the five test articles. More experience and training with these contexts might improve the intermarker consistency.

Table 4 shows the results of the intermarker consistency study for didactic case reports. The subjects, three nurses and two preclinical medical students, achieved moderate or closer agreement for 67 percent of the contexts in this model. Once again, several contexts that were rarely assigned to the test collection (i.e., *Differential diagnosis – management implications* and *Differential diagnosis –Prognostic implications*) had low kappa values. Low frequency in the collection cannot completely explain this inconsistency because almost perfect agreement was observed for the *Acknowledgments* context, which had the lowest frequency. Typically, only a few sentences

Table 3. Intermarker consistency for clinical research articles

Context name	Kappa	Interpretation	Frequency
Clinical research article			
Title	1.00	Almost perfect	0.006
Authors	0.96	Almost perfect	0.026
Background	0.79	Substantial	0.129
Objective	0.80	Substantial	0.019
Methods			
Study type	0.64	Substantial	0.006
Study setting	0.55	Moderate	0.004
Study-population methodology			
Eligibility/selection	0.70	Substantial	0.044
Exclusion/withdrawal	0.55	Moderate	0.017
Study-Group Assignment	0.00	Slight	<0.001
Experimental parameters			
Risk/Prognostic factor assessment	0.26	Fair	0.011
Intervention evaluated	0.62	Substantial	0.039
Concomitant interventions	0.04	Slight	0.003
Diagnostic test evaluated	0.68	Substantial	0.060
Gold standard	0.51	Moderate	0.014
Outcome measures/endpoints	0.76	Substantial	0.075
Consent/ethics procedures	0.75	Substantial	0.005
Statistical methods	0.88	Almost perfect	0.019
Results			
Experimental findings	0.79	Substantial	0.234
Adverse effects	0.51	Moderate	0.012
Conclusions	0.54	Moderate	0.068
Limitations/biases	0.50	Moderate	0.029
Future work	0.44	Moderate	0.006
Acknowledgments/collaborators	0.92	Almost perfect	0.008
References	1.00	Almost perfect	0.184

Table 4. Intermarker consistency for case reports

Context name	Kappa	Interpretation	Frequency
Case report			
Title	1.00	Almost perfect	0.008
Authors	1.00	Almost perfect	0.017
Discussants	0.98	Almost perfect	0.039
Case presentation			
Case findings	0.79	Substantial	0.413
Case management	0.38	Fair	0.032
Case discussion			
Differential diagnosis			
Ddx – diseases in differential diagnosis	0.70	Substantial	0.206
Ddx – Diagnostic test implications	0.12	Slight	0.017
Ddx – Management implications	0.27	Fair	0.007
Ddx – Prognostic implications	0.28	Fair	0.004
Definitive/presumed diagnosis			
Dx – Definitive/presumed diagnosis	0.54	Moderate	0.085
Dx – Diagnostic test implications	0.53	Moderate	0.025
Dx – Management implications	0.52	Moderate	0.029
Dx – Prognostic implications	0.38	Fair	0.010
Acknowledgments	1.00	Almost perfect	0.002
References	1.00	Almost perfect	0.121

in an article are associated with the *Acknowledgments* context. However, authors routinely acknowledge their colleagues and sources of funding in medical publications. Thus, its consistent presence in case reports might have provided the necessary experience in identifying sentences associated with this context.

Contexts with similar names, such as the contexts *Differential diagnosis – prognostic implications* and *Definitive/presumed diagnosis – prognostic implications* were also common sources of disagreement. Confusion about these distinctions or careless markup may explain the low kappa values for several of these parallel constructs.

Intermarker consistency for review articles was substantially less than the agreement for the other two context models. Two medical school graduates and three physicians trained in internal medicine participated in the study. Table 5 shows the results of this experiment. The kappa values for several compound contexts that did not occur in any articles were undefined. Although the overall kappa value for the remaining contexts was 52 percent, the context markers achieved moderate or closer agreement for only 36 percent of all contexts and 38 percent of non-compound contexts in the model. This experiment demonstrates the importance of both individual and overall kappa coefficients. The overall agreement in context markup was moderate, but the model had significant deficiencies that individual kappa values identified.

The most frequent sources of disagreement were the components of the two compound contexts, *Reviewed topics* and *Relevant topics*. We intended that these categories of information would distinguish topics that were the focus of a review from the secondary issues. The clinicians in our experiment felt that the differences between these categories were poorly defined and difficult to appreciate. The separation of *Relevant* and *Reviewed* information is probably an artificial distinction. The information that is "important" in a review depends on the interests and perspectives of the reader.

In designing the review context model, we characterized the features of a diverse set of publications, including summaries of medical knowledge and meta-analyses [15]. The semantic differences in these types of articles resulted in a context model that inadequately represented both publications and confused the context markers. Meta-analyses apply rigorous scientific methods in selecting clinical studies and in analyzing their data, and this information constitutes an important part of the article [16]. These publications share many features with clinical research articles. Authors of didactic or qualitative reviews may not employ systematic techniques for examining a topic, and the styles for writing these articles differ significantly across authors and journals. A single context model may not adequately characterize all forms of review articles.

5 Discussion

We have constructed a document representation that captures the contexts in which index terms occur. Through studies of intermarker consistency, we have demonstrated that clinicians can easily learn and apply context models for clinical research articles and case reports. The substantial agreement of context markup suggests that these models provide an effective indexing scheme for information retrieval. Since the context

Table 5. Intermarker consistency for review articles

Context name	Kappa	Interpretation	Frequency
Review article	0.00	Slight	< 0.001
Title	0.95	Almost perfect	0.004
Authors	0.89	Almost perfect	0.013
Objective	0.63	Substantial	0.007
Background	0.85	Almost perfect	0.025
Reviewed topic(s)	Undefined		0.000
Reviewed population or diseases	0.06	Slight	0.060
Reviewed interventions	0.20	Slight	0.056
Reviewed risk factors/exposures	0.05	Slight	0.019
Reviewed tests	0.15	Slight	0.004
Reviewed outcomes	0.12	Slight	0.039
Reviewed adverse effects/complications	0.27	Fair	0.139
Relevant topic(s)	Undefined		0.000
Relevant population or diseases	0.19	Slight	0.012
Relevant interventions	0.00	Slight	0.008
Relevant risk factors/exposures	0.02	Slight	0.010
Relevant tests	0.09	Slight	0.009
Relevant outcomes	0.00	Slight	< 0.001
Relevant adverse effects/complications	−0.0009	Poor	0.003
Review methodology	0.00	Slight	0.001
Data source selection procedures	0.59	Moderate	0.021
Data source exclusion procedures	0.68	Substantial	0.002
Data extraction/statistical methods	0.61	Substantial	0.034
Review results	Undefined		0.000
Review data	0.30	Fair	0.180
Sources of variation	0.07	Slight	0.045
Biases/limitations	0.28	Fair	0.072
Conclusions	0.61	Substantial	0.061
Future work	0.25	Fair	< 0.001
Acknowledgments	0.82	Almost perfect	0.005
References	1.00	Almost perfect	0.363

markers were health care professionals and trainees, our results suggest that the users of retrieval systems could readily assimilate the context models as a basis for searching. With our previous findings that these structures can improve the precision of searches in the medical literature [28], these results support the use of a contextual structure in designing full-text digital libraries.

Context modeling may not be appropriate for all types of publications. Defining a semantic structure for articles that exhibit broad stylistic variations (e.g., didactic reviews), may be impossible. For such publications, a context-based representation may be unnecessary. Reviews are often well characterized by their titles, and traditional techniques for searching may be sufficient.

Our research has several limitations. First, we evaluated context models using the literature of a single medical specialty. The journal articles in our experiments predominately addressed issues of internal medicine. However, both the *Journal of the American Medical Association* and the *New England Journal of Medicine* publish research from a variety of medical disciplines. Our test collection contained articles from surgery, dermatology, and neurology. Although we expect that the publications of other medical disciplines share common themes, our conclusions should not be applied in these fields without additional studies that replicate our findings. Second, the context markers were willing volunteers, and they were aware that their performance was being evaluated. Thus, the substantial intermarker consistency for clinical research articles and case reports may be artificially high due to both self-selection and a Hawthorne effect. Nonetheless, these biases probably do not account fully for intermarker consistency that was equal to or greater than the agreement of professional indexers. Finally, the kappa coefficient assumes that all disagreements about context assignments are equal. In practice, certain inconsistencies in markup may be more significant than others. For example, if one context marker assigned a sentence to the *Eligibility/selection* context, the selection of *Study-group assignment* by another context marker seems a less inconsistent than the choice of *Exclusion/withdrawal*. Although researchers can represent different levels of agreement by weighting the kappa coefficient, weighting schemes are arbitrary and detract from the interpretation of this measure [22]. Because we did not acknowledge partial agreement in our evaluation, our results may underestimate the degree to which context markers consistently applied the models.

We suggest that context-based representations are valuable for bodies of literature that exhibit characteristic presentation styles and that discuss topics with considerable overlap in subject matter. Stereotypic semantic structures are essential for defining useful context models. Texts that describe the literature [35] or recommendations for authors [2, 14] usually outline the recurring themes in a publication. Frequently used section headings also suggest representative types of information, even if a specific heading does not always accurately characterize the contents of a section. These resources are useful starting points in developing context models.

Context models provide leverage in searching collections with significant overlap in subject areas. Medical publications traditionally describe the prior work that motivated or influenced the current research. In addition, medical publications may address diverse aspects of the same medical topic. Such redundancy can undermine the performance of information-retrieval systems that rely on term-based representations. Many scientific domains have bodies of literature with features similar to the literature of the field of medicine. Information-retrieval systems for these disciplines are likely to benefit from context-based representations.

6 Conclusions and future research

We are currently developing context models for other medical publications. Meta-analyses exhibit a distinct and characteristic structure that is amenable to representation in a context model [34]. Didactic reviews share features with other educational resources such as textbooks and board-review materials. We are revising our original model for review articles to capture these common themes. Context models represent single classes of publications, but most clinicians want to search across collections of diverse resources. A practical context-based retrieval system must provide searching in a variety of publications through a single, comprehensive interface. To develop such a tool, we are exploring the relationships among context models and their associations with clinical information needs.

In conclusion, we have presented a robust document representation that explicitly models context in medical publications. Our evaluation of intermarker consistency demonstrates that clinicians can easily comprehend and apply the context models for two types of publications with minimal training. Other experiments have demonstrated that this representation can improve the precision of searches in full-text collections of the medical literature [28]. These findings support the use of context models in document representations for full-text databases. This technique may be applicable to other domains that have characteristic styles of presentation and to publications that describe diverse aspects of similar topics. For these bodies of literature, the contextual structure can augment term representations by providing meaning for words from the text. The contexts facilitate interpretation of index terms, and thus, serve as the basis for precise retrieval. Designers of digital libraries must consider the underlying representation of their content to provide efficient retrieval and display of diverse information resources.

Acknowledgements. We thank Leslie Lenert, Thomas Rindfleisch, Michael Walker, and Terry Winograd for their contributions to this work. The National Library of Medicine supported this project through grant LM-07033, and the Section on Medical Informatics provided computing facilities with grant LM-05208. We are greatly indebted to the medical students, residents, physicians, and nurses who donated their time to participating in these experiments.

References

1. Ad Hoc Working Group for Critical Appraisal of the Medical Literature. A proposal for more informative abstracts of clinical articles. *Annals of Internal Medicine*, 106:598–604, 1987
2. Asilomar Working Group on Recommendations for Reporting of Clinical Trials in the Biomedical Literature. Checklist of information for inclusion in reports of clinical trials. *Annals of Internal Medicine*, 124(8):741–43, 1996
3. Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960
4. Fleiss, J. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–82, 1971
5. Froom, P., Froom, J. Deficiencies in structured medical abstracts. *Journal of Clinical Epidemiology*, 46(7):591–94, 1993
6. Funk, M. E., Reid, C. A. Indexing consistency in MEDLINE. *Bulletin of the Medical Library Association*, 71(2):176–83, 1983
7. Glowniak, J. V. Medical resources on the internet. *Annals of Internal Medicine*, 123(2):123–31, 1995
8. Heller, M. B. Structured abstracts: a modest dissent. *Journal of Clinical Epidemiology*, 44(8):739–40, 1991
9. Hersh, W. R., Hickam, D. H. A comparison of retrieval effectiveness for three methods of indexing medical literature. *American Journal of the Medical Sciences*, 303(5):292–300, 1992
10. Hersh, W. R., Hickam, D. H., Haynes, R. B., McKibbon, K. A. A performance and failure analysis of SAPHIRE with a MEDLINE test collection. *Journal of the American Medical Informatics Association*, 1(1):51–60, 1994
11. Hogan, R. Medical CD-ROMs. *Journal of the American Medical Association*, 270(13):1613–15, 1993
12. Humphrey, S.M. Indexing biomedical documents: from thesaural to knowledge-based retrieval systems. *Artificial Intelligence in Medicine*, 4:343–71, 1992
13. Huth, E. J. Structured abstracts for papers reporting clinical trials. *Annals of Internal Medicine*, 106(4):626–27, 1987
14. International Committee of Medical Journal Editors. *Uniform requirements for manuscripts submitted to biomedical journals*. 1994
15. Kass, E. H. Reviewing reviews. In K. S. Warren (Ed.), *Coping with the biomedical literature* (pp. 79–91). New York: Praeger Publishers, 1981
16. L'Abbe, K. A., Detsky, A. S., O'Rourke, K. Meta-analysis in clinical research. *Annals of Internal Medicine*, 107(2):224–33, 1987
17. Lacroix, E. M., Backus, J. E., Lyon, B. J. Service providers and users discover the Internet. *Bulletin of the Medical Library Association*, 82(4):412–18, 1994
18. Lancaster, F. W. *Evaluation of the MEDLARS demand search service*. U.S. Dept. of Health, Education, and Welfare, Public Health Service, 1968
19. Landis, J. R., Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174, 1977
20. Leonard, L. E. *Inter-indexer consistency and retrieval effectiveness: measurement of relationships*. Ph.D. Thesis, University of Illinois, Champaign, IL, 1975
21. Lindberg, D. A. B., Humphreys, B. L., McCray, A. T. The unified medical language system. *Methods of Information in Medicine*, 32(4):281–91, 1993
22. Maclure, M., Willett, W. Misinterpretation and misuse of the kappa statistic. *American Journal of Epidemiology*, 126(2):161–69, 1987
23. McKibbon, K. A., Haynes, R. B., Dilks, C. J. W., Ramsden, M. F., Ryan, N. C., Baker, L., Flemming, T., Fitzgerald, D. How good are clinical MEDLINE searches? A comparative study of clinical end-user and librarian searches. *Computers and Biomedical Research*, 23:583–93, 1990
24. Mendelson, D. N., Levinson, J., Gaylin, D. S. The anatomy of online information for physicians. *Canadian Medical Association Journal*, 155(6):665–74, 1996
25. Mulrow, C. D., Thacker, S. B., Pugh, J. A. A proposal for more informative abstracts of review articles. *Annals of Internal Medicine*, 108:613–15, 1988
26. National Library of Medicine. *Medical subject headings – annotated alphabetic list 1995*. Bethesda, MD: US Department of Health and Human Services, Public Health Service, National Institutes of Health, 1994
27. National Library Of Medicine. *Medical subject headings – annotated alphabetic list 1996*. Bethesda, MD: US Department of Health and Human Services, Public Health Service, National Institutes Of Health, 1995
28. Purcell, G. P. *Contextual document models for searching the clinical literature*. Ph.D. Thesis, Section on Medical Informatics, Stanford University, Stanford, CA, 1996
29. Purcell, G. P. Task-specific extracts for using the medical literature. Presented at the Northern California Medical Library Group/Medical Library Group of Southern California Joint Meeting, Berkeley, CA, January 1997
30. Relman, A. S. Journals. In K. S. Warren (Ed.), *Coping with the biomedical literature* (pp. 67–78). New York: Praeger Publishers, 1981
31. Salton, G., McGill, M. J. *Introduction to modern information retrieval*. New York: McGraw-Hill, 1983
32. Schuyler, P. L., Hole, W. T., Tuttle, M. S., Sherertz, D. D. The UMLS Metathesaurus: representing different views of biomedical concepts. *Bulletin of the Medical Library Association*, 81(2):217–22, 1993
33. Sievert, M. C., McKinin, E. J., Johnson, E. D. Full-text databases in medicine. *Journal of the American Society for Information Science*, 46(10):748–54, 1995
34. Thacker, S. B. Meta-analysis: a quantitative approach to research integration. *Journal of the American Medical Association*, 259(11):1685–9, 1988
35. Warren, K. S. *Coping with the biomedical literature*. New York: Praeger Publishers, 1981