Research Paper ■

Automatically Identifying Health Outcome Information in MEDLINE Records

DINA DEMNER-FUSHMAN, MD, PHD, BARBARA FEW, RN, MSN, MSI, SUSAN E. HAUSER, PHD, GEORGE THOMA, PHD

Abstract **Objective:** Understanding the effect of a given intervention on the patient's health outcome is one of the key elements in providing optimal patient care. This study presents a methodology for automatic identification of outcomes-related information in medical text and evaluates its potential in satisfying clinical information needs related to health care outcomes.

Design: An annotation scheme based on an evidence-based medicine model for critical appraisal of evidence was developed and used to annotate 633 MEDLINE citations. Textual, structural, and meta-information features essential to outcome identification were learned from the created collection and used to develop an automatic system. Accuracy of automatic outcome identification was assessed in an intrinsic evaluation and in an extrinsic evaluation, in which ranking of MEDLINE search results obtained using PubMed Clinical Queries relied on identified outcome statements.

Measurements: The accuracy and positive predictive value of outcome identification were calculated. Effectiveness of the outcome-based ranking was measured using mean average precision and precision at rank 10.

Results: Automatic outcome identification achieved 88% to 93% accuracy. The positive predictive value of individual sentences identified as outcomes ranged from 30% to 37%. Outcome-based ranking improved retrieval accuracy, tripling mean average precision and achieving 389% improvement in precision at rank 10.

Conclusion: Preliminary results in outcome-based document ranking show potential validity of the evidence-based medicine–model approach in timely delivery of information critical to clinical decision support at the point of service.

■ **J Am Med Inform Assoc.** 2006;13:52–60. DOI 10.1197/jamia.M1911.

Introduction

Objective

To better satisfy clinical information needs related to health care outcomes, we set out to develop a methodology for automatic identification of outcomes-related information in medical text. If outcomes information can be identified, it may serve as one of the indicators of clinical orientation of the text, and be presented to clinicians as a starting point for decision making. In this article, we propose outcome identification as a method for guiding medical domain-specific search for relevant information.

Motivation

Scientific literature is a well-established source of information for professionals who need to stay current in their fields. The amount of information published in many fields makes the exhaustive coverage of information not only impractical, but practically impossible.[1] Not surprisingly, automatic methods that help these professionals find relevant information without evaluating every publication remain the focus of research in many text-related tasks. In the medical domain, in addition to improvements in information access, automatic methods have the potential to enable evidence-based practice.[2]

To provide optimal patient care, a clinician needs just-in-time access to the best available evidence in the context of the patient's individual condition. One of the key elements in using current best evidence while making decisions about the care of individual patients is an understanding of the effect of a given intervention on the patient's health outcome and quality of life.[3] However, the well-documented difficulty clinicians have dealing with the sheer volume of medical literature[4–7] may prevent them from accessing knowledge about health outcomes at the time the information is required. When information about health care outcomes is not known, a physician's ability to provide optimal patient care may be compromised.

The task of identifying outcomes-related information is twofold: determine the minimal amount of the text required to understand the health outcome implications of the research paper, and then identify these text units in clinical texts. The latter essentially can be viewed as text classification. We addressed the first part of the task, studying how much information is required and what level of granularity is sufficient to estimate potential clinical validity of a publication using health outcome information.[8] These experiments lead us to select passages in article abstracts as units appropriate for the outcome identification task. Our approach to the second

Affiliation of the authors: Lister Hill National Center for Biomedical Communications, Communications Engineering Branch, National Library of Medicine, National Institutes of Health, Bethesda, MD.

Correspondence and reprints: Dina Demner-Fushman, MD, PhD, Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD 20894; e-mail: <ddemner@mail.nih.gov>.

Received for review: 07/15/05; accepted for publication: 09/15/05.

part of the task as a classification problem utilizes well-known text categorization techniques as the baseline,[9,10] and improves the initial result using an ensemble of classifiers.

The classification task and evaluation of the classification results require a "truth set"—a collection of documents with passages annotated as outcome statements. Unlike medical literature classification tasks that can use existing resources,[11] our task required creation of the test collection of annotated article abstracts. We selected MEDLINE as the most likely resource to be used by a practitioner. Our choice is based on results of studies that show MEDLINE to be a primary source for questions pertaining to treatment,[12] and the database most frequently searched by clinicians.[13] Moreover, in a study by Hersh et al.,[14] all users that used MEDLINE searches with subsequent access to the full text of the article improved their ability to answer clinical questions.

In the remainder of this article, we first review related work, then describe the creation of the test collection and the outcome identification methodology, next we review and discuss the results of our experiments, and finally we conclude with a summary of our contribution and a discussion of future directions.

Background

To our knowledge, identification of clinical outcome statements has not been previously approached as a text classification task. Text classification amounts to labeling a given text, most frequently a news article or a MEDLINE citation[a] with one or more predefined categories that reflect the topicality, genre, authorship, etc. (See Sebastiani[10] for a thorough review of modern methods widely applied in text classification.) It remains to be seen whether the same methods can be successfully applied to classification of text passages that constitute outcome statements. Preliminary results in classification of MEDLINE citation sentences as belonging to one of the sections of structured abstracts[11] or as speculative statements[15] are encouraging. However Teufel and Moens[16] had to use nontextual features as, for example, absolute or relative location of a sentence and section structure, in addition to the text of the sentence in assignment of rhetorical status to sentences for the purpose of summarization of scientific articles. Rhetorical roles of the sentences reflect the domain-dependent structure of the documents.[17] Medical articles combine the generic structure of scientific articles[18] with the specific domain elements. Purcell et al.[19] captured the complex structure of medical articles in three hierarchical context models for medical document representation, and identified a number of outcome-oriented elements including experimental findings, reviewed outcomes, and relevant outcomes. In view of these findings, we incorporate the discourse structure of the abstracts of medical articles as one of the classifiers in our ensemble.

Identification of outcome statements in the so-called secondary sources, i.e., in the key treatment recommendations on important clinical topics compiled by experts, was undertaken by Niu and Hirst.[20] Their approach starts with the identification of the cue words indicative of outcome statements in a sentence. Once the cue word is identified, the boundaries of the statement are determined using rules based on the part of speech of the cue word. It has not been determined whether the resulting statements can serve as elementary units sufficient for understanding the health outcome implications in medical articles and their abstracts. It has been observed that providing users with information within its context is preferable to mere presentation of the fact.[21] This observation is in concert with the success of the systems that provide summaries of literature personalized to an individual patient history[22,23] or that formulate queries using a patient's information.[24]

Classification methods have been successfully applied to medical text in the past. Ensembles of classifiers were particularly successful, and consistently outperformed single classifiers in medical text categorization, e.g., in assigning ICD9 codes to patient discharge summaries,[25] and in assigning Heart Disease categories to MEDLINE abstracts.[26,27] An explanation of better performance of the ensemble methods is given in the overview of methods for combining classifiers with the goal of improving classification results.[28] Once the decision to combine classifiers is made, some consideration should be given to selecting a promising method. We implemented the stacking method developed and tested on several datasets from the repository of machine learning[29] because it is recommended for the cases where base classifiers are disparate in nature. The text classification method closest to the stacking method used in our experiments is a probabilistic method of combining classifiers shown to be successful on news-wire text.[30]

Methods

The supervised machine learning approach to outcome classification task and evaluation of the classification results depend on the availability of test collections of documents. We developed a collection of 633 MEDLINE citations, in 592 of which we identified and annotated passages containing outcome statements. The collection was divided into a training set used to develop an automatic outcome identification methodology, and a test set used for intrinsic evaluation of the automatic outcome classification. The intrinsic evaluation pursued two goals. First, establish the validity of the developed outcome identification methods. Second, verify that the methods are generic enough to be applied independent of a clinical scenario, potentially generating four major types of questions within the context of major clinical tasks: etiology, diagnosis, therapy, and prognosis.[3] In addition to the intrinsic evaluation, we conducted an extrinsic evaluation of our method in an information retrieval task, in which PubMed retrieval results were ranked based on the results of the final selection of the outcome passage.

Outcome Identification and the Intrinsic Evaluation

The Test Collection for Machine Learning and Intrinsic Evaluation

Our test collection consists of five sets of MEDLINE citations created emulating different types of user behaviors. The search strategies used to query PubMed are presented in Appendix 1. Our goal was to annotate succinct patient health outcome statements in the abstracts of the citations. We used

[a]In bibliographic databases such as MEDLINE, the base record is a citation to an article, book, or other document. The citation includes the title and the abstract of the item, subject headings, author(s), publication type(s), date of publication, language, etc.

the MeSH scope notes to define the outcome as "…the results or consequences of management and procedures used in combating disease…," and the Problem-Intervention-Comparison-Outcome (PICO) framework[31] developed to help physicians create successful strategies for searching medical literature to identify outcomes as a components of the PICO framework. The initial set (Set 1) was generated using three disease categories: rheumatoid arthritis, migraine, and breast cancer. Core clinical journals were searched for a 1-year time period using search strategies developed by the HEDGES Study Team.[32,33] A set of 356 retrieved citations was narrowed to 275 citations after eliminating studies (1) evaluating patient questionnaires, rating scales, and the like; (2) outside of the disease categories; (3) without abstracts; and (4) not dealing with treatment outcomes. The registered nurse (RN) author, a clinical nurse specialist with more than 20 years of experience, annotated the 275 abstracts identifying sentences containing health outcomes. On average, 2.25 sentences per abstract were annotated as outcome statements.

The same search strategies were then applied to obtain the second set of citations (Set 2) focusing on three chronic diseases: pulmonary tuberculosis, renal hypertension, and asthma. The RN annotated on average 1.9 sentences per abstract as outcome statements in this set. The second annotator, a medical student, on average annotated 4.3 sentences per abstract as outcome statements; 83% of the statements identified as outcomes by the RN were also marked as such by the student. Because of this large difference in the size of the annotated passages, agreement between the annotators was only fair (kappa = 0.42). Analysis of the disagreements showed that the medical student tended to include disease-oriented outcomes and statistical information in support of the outcome in addition to the patient outcome statements (Fig. 1).

To provide guidance for annotators and promote consistency among them, we reviewed our annotation scheme and extended it to seven categories, separating the outcome statement from its supporting text (Table 1). Our final scheme is very similar to the PP-ICONS approach for identification of valid or relevant articles[34] developed in parallel and unknown to us at the time our annotation and scheme development took place.

Table 1 ■ Scheme for Annotation of Clinically Relevant Elements in MEDLINE Citations

Tag	Definition
Background	Material that informs and may place the current study in perspective, e.g., work that preceded the current; information about disease prevalence, etc.
Population	The group of individual persons, objects, or items comprising the study's sample, or from which the sample was taken for statistical measurement
Intervention	The act of interfering with a condition to modify it or with a process to change its course (includes prevention)
Statistics	Data collected about the results of the intervention demonstrating its effect
Outcome	The sentence(s) that best summarizes the consequences of an intervention
Supposition	An assumption or conclusion that goes beyond the evidence presented in an abstract
Other	Any sentence not falling into one of the other categories and presumed to provide little help with clinical decision making.

A

Treatment with losartan in patients with type 2 diabetes and nephropathy not only reduced the incidence of ESRD, but also resulted in substantial cost savings.

B

…losartan and CT compared with placebo and CT reduced the number of days with ESRD by 33.6 per patient over 3.5 years (P = 0.004, 95% CI 10.9-56.3). This reduction in ESRD days resulted in a decrease in cost associated with ESRD of 5144 US dollars per patient (P = 0.003, 95% CI 1701 to 8587 US dollars). After accounting for the cost of losartan, the reduction in ESRD days resulted in a net savings of 3522 US dollars per patient over 3.5 years (P = 0.041, 143 to 6900 US dollars). CONCLUSIONS: Treatment with losartan in patients with type 2 diabetes and nephropathy not only reduced the incidence of ESRD, but also resulted in substantial cost savings.

Figure 1. (**A**) An outcome statement identified by the RN. (**B**) An outcome statement identified by the medical student.

Two citation sets (Set 3 and Set 4) were created using "typical" queries, i.e., without advanced search criteria available in PubMed using only *language*, *human*, and *only items with abstracts* limits. The 50 most recent abstracts were selected for annotation from each of the sets, resulting in 100 citations annotated using the new scheme. Sixty-six of these citations contain outcome statements. The fifth set (Set 5) was previously annotated in a study of granularity level necessary for estimating the clinical validity of MEDLINE citations.[8] Set 5 provides outcome statements annotated in 128 citations that were indexed for MEDLINE as containing treatment outcomes. The five subsets of the test collection are shown in Table 2.

Before using the outcome statements as the reference standard in the intrinsic evaluation, raters reconciled their differences in annotation and came to consensus for sets 4 and 5. The intersection of two annotations constitutes true positives for Set 2, and sentences marked as an outcome by the majority of the experts are considered to be true positives in Set 3.

Automatic Outcome Identification

Classifiers for automatic outcome identification operate on sentences of the abstracts of MEDLINE citations and estimate

Table 2 ■ Five Sets of MEDLINE Citations That Form the Test Collection

Set	Search Topic	Annotators	Citations	With Outcome
1	Rheumatoid arthritis, migraine, breast cancer	RN1	275	275
2	Exercise-induced asthma, renal hypertension, tuberculosis	RN1, medical student	123	123
3	Immunization	RN1, RN2, MD, PhD	50	33
4	Diabetes	RN1, RN2, MD, PhD	50	33
5	Treatment Outcome [mh]	MD, PhD	135	128
Total			633	592

the likelihood or probability that each sentence belongs to an outcome statement. In this section, we first present the final split of the collection into the test and training sets based on our preliminary experiments with the readily available machine learning toolkits[35,36] and traditional random splits of the collection.[37] Then we describe how each classifier works, the choice of classifiers, and features used in classification. The last part of this section presents methods used to combine classifiers.

From our collection of 592 citations with annotated outcome statements, we chose to use the 275 abstracts of Set 1 as our training set, leaving 317 citations with outcome statements and 41 without (358 total) for testing. The choice was made after experimenting with the sizes of the sets following recommendations that the test set size should be 5% to 10% of the collection size.[37] We conducted the preliminary experiments using the WEKA toolkit[35] as follows: 10 iterations of randomly selecting from 633 citations and setting aside 60 as the test set and another 60 citations as the verification set and using the rest for training. In these experiments, the relatively large training set did not improve the classification results over the results obtained using Set 1 for training and Set 2 for testing. Furthermore, the small size of the test set prevented testing the system's performance for each of the four clinical tasks. In the subsequent experiments we settled on the 275 citations as the training set sufficient to maintain the performance achieved in the preliminary experiments.

In these experiments, a Naïve Bayes classifier outperformed both a linear SVM and a decision-tree classifier in identifying outcome statements, and was selected as the baseline classifier for further experiments. In additional preliminary experiments, we used the state-of-art Naïve Bayes classifier provided with the MALLET toolkit.[36] This Naïve Bayes classifier achieved 100% recall and 27% precision, prompting us to create a coordinated ensemble of classifiers, i.e., train complementary classifiers, then classify sentences in each citation using (1) linear interpolation with ad hoc weights assigned based on intuition and (2) a weighted sum of the classifiers combined in optimum way using stacking.[29] The six base classifiers in this ensemble are: a rule-based classifier, a Naïve Bayes classifier, an n-gram–based classifier, a position classifier, a document length classifier, and a semantic classifier. The rules for the rule-based classifier were created manually by the RN author. The remaining classifiers were trained on the 275 citations from the annotated collection described above.

Outcome identification starts with a classification of each sentence in the abstract as an outcome statement or not. Each of the base classifiers described below generates either a likelihood estimate, or a probability that the sentence belongs to the outcome. The rule-based, Naïve Bayes, and n-gram–based classifiers treat each sentence disregarding the context of the abstract. The position classifier and the semantic classifier use the abstract structure and context, and the document length classifier operates solely on the number of sentences in the abstract.

The rule-based classifier estimates likelihood of the sentence to be an outcome based on cue phrases such as "significantly greater," "well tolerated," and "adverse events." The strength of evidence provided by cue phrases is measured by the ratio of the cumulative score for found phrases to maximal possible score. For example, the following sentence: "The dropout rate due to adverse events was 12.4% in the moxonidine and 9.8% in the nitrendipine group" is segmented into eight phrases during MetaMap[38] processing, so the maximal possible score is set to 8, and the two phrases "dropout rate" and "adverse events" contribute 1 point each to the cumulative score, which results in a likelihood estimate of 0.25 for the sentence.

The Naïve Bayes classifier treats each sentence as a bag of words and generates the probability of the sentence to be an outcome statement, rather than a binary decision with respect to the class of the sentence being an outcome or not.

The n-gram–based classifier generates the probability in a manner different from the Naïve Bayes classifier: whereas the probability assigned by the Naïve Bayes classifier is based on probabilities of all words encountered during training, the n-gram–based classifier uses only features that are strong positive predictors of outcomes. These features were selected as uni- and bi-grams by first identifying the most informative features using information gain measure,[39] then selecting only positive outcome predictors using odds ratio,[40] and finally by a manual revision by the RN author (during which the topic-specific terms, such as rheumatoid arthritis, one of the three diseases used to retrieve the training documents, were removed from the feature set to ensure generality of the features). As an example, consider the terms "superior" and "placebo controlled." Both have a high information gain value, but "superior" also has a high positive odds ratio value and is selected as a feature for the n-gram classifier, as opposed to "placebo controlled" that has a high negative odds ratio, and is therefore discarded.

The position classifier is based on the discourse structure of the abstract and the relative position of the sentence in the abstract. As can be seen in Figure 2, the likelihood estimate that a sentence contains an outcome statement is very high for the last 3 sentences of the abstract. This is also true for the sentences in the results and the conclusions sections of the structured abstracts. Of the 275 citations used for training, 22 (2.5%) were not structured. In the rest, the outcome statements were found in conclusions in 63.6% of the structured abstracts, in the results section of another 36%, and in the interventions section of 1 abstract.

The document length classifier returns a smoothed probability that a document of given length (in the number of sentences) contains an outcome statement. For example, the probability that a 3-sentence-long abstract contains an outcome statement

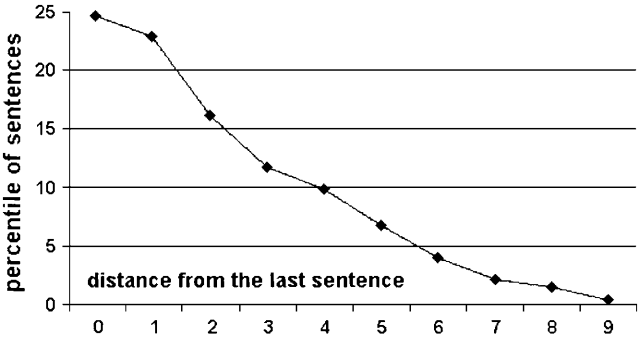

Figure 2. Positions of outcome statements in 275 training abstracts.

is 0.2, and the probability to find an outcome statement in an 11- to 14-sentence-long abstract is 0.95. Implementation of this classifier is motivated by the observed difference between the lengths of abstracts in which the outcome statements were found, and were not found. The average length of the former is 11.7 sentences, whereas the length of the latter is 7.95 sentences on average. We use the document length classifier with the understanding that it is meaningless when identifying sentences containing outcomes within the document, but will boost our confidence when selecting documents most likely to contain outcome statements.

The semantic classifier generates the maximum likelihood estimate of a given sentence being an outcome statement based on the presence of UMLS concepts belonging to semantic groups highly associated with outcomes, such as therapeutic procedure or pharmacologic substance. Identification of the semantic groups is based on mappings of the semantic types to the groups.[41] The underlying identification of the UMLS concepts and semantic types associated with these concepts is achieved using MetaMap.[38] The semantic classifier is global, i.e., it takes into consideration the previously seen content of an abstract temporarily stored during its sequential processing. For example, if the problem and interventions identified in a sentence using MetaMap processing correspond to those named in the title and the objectives section of the abstract, the likelihood estimate that the sentence is an outcome statement increases.

The probabilities and likelihood estimates of being an outcome statement (assigned to the sentence by the base classifiers) are then combined by the meta-classifier using ad hoc weights selected based on our intuitions about the prediction of the base classifier. We also experimented with optimum combination of weights using confidence values generated by the base classifiers and stacking—the version of least squares linear regression adapted for classification tasks.[29] This multiple linear regression (MLR) meta-classifier, which has been shown to outperform other methods of combining classifiers, can be described by the following equation:

$$MLR(x) = \sum_{k}^{K} \alpha_k P_k(x)$$

$P_k(x)$ is the probability that sentence x belongs to an outcome statement, as determined by classifier k (for classifiers that do not return actual probabilities, we use likelihood estimates). To predict the class of a sentence, the probabilities generated by K classifiers are combined using the coefficients (α_0, …, α_k). The coefficients' values are determined in the training stage as follows: probabilities predicted by base classifiers for each sentence are represented as a KN matrix A, where N is the number of sentences in the training set, and K is the number of classifiers. The reference set class assignments for each sentence are stored in a vector b, and the coefficients' values are found by calculating the vector α that minimizes $\|A\alpha - b\|$. The coefficients were found using singular value decomposition (SVD), as provided in the JAMA basic linear algebra package released by NIST.

Intrinsic Evaluation Methodology

We evaluated automatic outcome identification using 2 different ways to combine base classifiers–ad hoc and stacking–for each of the 4 main physician's tasks: etiology, diagnosis, therapy, and prognosis.[3,33,42] Abstracts in the test set fell into task categories as follows: 37 abstracts pertain to etiology, 57 to diagnosis, 153 abstracts to therapy, and 111 to prognosis. Citations were assigned to categories based on search strategies developed for searching MEDLINE[32] and Users' Guides to Evidence-based Medicine[43] (EBM) as follows: citations containing *etiology* and *cohort studies* in their major MeSH headings were assigned to the etiology category. Citations containing the *diagnosis* MeSH heading were assigned to the diagnosis task. Citations indexed with *therapy* and *therapeutic use* headings fell into the therapy category. Citations indexed with MeSH headings *follow-up studies*, *quality of life*, and *mortality* as major topics were assigned to the prognosis category.

The output of each of the outcome meta-classifiers for a single abstract is a list of sentences ranked in descending order by the confidence score assigned to the sentence. Based on the observation that annotators typically marked 2 to 3 sentences in each abstract as outcomes, we evaluated the performance of our meta-classifiers at cutoffs of 2 and 3 sentences. In addition we evaluated accuracy of selecting just 1 sentence, as a possible help in rapid assessment of a citation's relevance to a clinical task. Motivated by the general expectation that patient outcome statements are typically found in the conclusion section of a structured abstract and toward the end of an unstructured abstract, we compared our outcome classifiers to a baseline of returning 1, 2, or 3 final sentences of an abstract (baselines 1 through 3, respectively, in Table 3).

Extrinsic Evaluation of Automatic Outcome Identification

The ultimate measure of success of our outcome identification method is its performance in a real-life task. For a preliminary evaluation in a real-life information retrieval task, we carried out automatic outcome-based ranking of 1,312 MEDLINE citations.

The Test Collection for Extrinsic Evaluation

The 1,312 citations were retrieved to answer clinical inquiries for five disorders using the sensitive and therapy-oriented Clinical Query available in PubMed. A family practitioner provided relevance judgments for 40 citations retrieved for each of the disorders.[44]

Extrinsic Evaluation Methodology

For the purposes of this experiment, an outcome statement consisted of the three top-ranking sentences generated by the stacking outcome meta-classifier. The confidence score of the top-ranking sentence was assigned to the whole outcome statement. The ranking of the citations started with an initial filter that discarded potentially irrelevant citations if the disorder automatically identified in the outcome was not the same as in the query. The remaining citations were ranked according to the sum of the outcome confidence score and the score of the potential strength of evidence presented in the citation. The potential strength of evidence score was determined according to the principles defined in the Strength of Recommendations Taxonomy.[45] The strength of evidence score was based on MeSH headings, such as publication type; publication in a core clinical journal having therapy as a major topic; and recent publication.

We evaluated the efficiency of the outcome-based ranking of the citations using mean average precision, and precision at rank 10, two of the measures developed at NIST to evaluate

Table 3 ■ Percent of Correctly Identified Outcome Statements at Three Cutoff Levels for Each Major Clinical Task

Classifier	Baseline			Meta-Classifier					
				Rule-based			Stacking		
Amount / Task	1	2	3	1	2	3	1	2	3
Etiology	34.5%	63.6%	78.2%	47.4%	68.4%	82.5%	52.6%	73.7%	87.7%
Diagnosis	44.4%	72.2%	75.0%	56.8%	70.3%	78.4%	67.6%	78.4%	89.2%
Therapy	38.6%	74.0%	75.0%	49.0%	75.0%	95.0%	51.0%	77.0%	92.8%
Prognosis	49.5%	73.0%	84.7%	63.1%	75.7%	87.4%	60.4%	79.3%	89.2%

retrieval systems that return ranked lists of documents.[46] Mean average precision for multiple topics (five disorders in our case) has recall and precision components. It is the mean of the average precision scores for each of the topics. The average precision score for a single topic is computed by averaging the precision after each relevant document is retrieved. We computed the percent improvement in precision at rank 10 as the difference in mean of relevant documents in the top 10 after and before the outcome-based reranking divided by the mean of relevant documents in the top 10 before reranking.

Results

Intrinsic Evaluation

Upper Bound for Outcome Identification

The upper bound for outcome identification was established using the interannotator agreement (Cohen's kappa[47]) on a sentence-by-sentence basis, i.e., whether the experts agreed on each sentence annotated as an outcome by at least one of them. Table 4 presents kappa values for the subsets of the collection described in the methods section. Two registered nurses, 1 medical student, 1 PhD, and 1 MD participated in the annotation of the collection. The RNs, the PhD, and the MD also had an opportunity to reconcile the differences in annotation for sets 4 and 5, which permitted measuring the intra-annotator reliability. For intra-annotator reliability, the original judgments made by a reviewer were compared with the consensus annotation. The intra-annotator consistency in annotating outcomes in Set 4 was excellent for clinicians (kappa ranging from 0.93 to 0.99) and good for the PhD (kappa = 0.81). Overall, clinicians had fewer difficulties with annotation: the intra-annotator consistency on all elements presented in Table 1 was 0.8, 0.9, and 0.92 for clinicians; whereas it was 0.73 for the PhD. On the fifth set, the intra-annotator consistency was 0.84 for the PhD and 0.91 for the MD.

The performance of the system was compared with that of the reviewers by computing kappa between the system and the combined reference standard of the human annotations (kappa = 0.67).

Accuracy of the Automatic Outcome Identification

The results of outcome identification are shown in Table 3, where numbers 1 through 3 indicate the sentence cutoffs in selecting sentences with top scores assigned by the outcome classifiers. In the evaluation, the prediction of the outcome classifier was considered correct if the sentences it returned intersected with sentences annotated as outcomes in the reference standard. We selected this lenient evaluation because of the importance of pointing the physician in the right direction, even if the results are only partially relevant. To complete the picture, we present the positive predictive value (computed as the ratio of true positive outcome sentences identified by the system to the sum of the true and false positive outcome sentences) of the top three sentences output by the stacking classifier in Table 5.

Extrinsic Evaluation: Outcome-based Ranking of Retrieval Results

Using mean average precision (map)—a measure frequently used in official NIST evaluations—we obtained a three-fold

Table 4 ■ Interannotator Agreement for Outcome Only Annotation (Outcome) and Annotation of All Clinically Relevant Elements (All Elements) in Outcome Identification

Set	Annotators	Annotation	All Annotators	Clinicians	Best Pairwise
2	RN1, MS	Outcome	0.42	0.42	0.42
3	RN1, RN2, PhD, MD	All elements	0.65	0.63	0.75
3	RN1, RN2, PhD, MD	Outcome	0.81	0.8	0.98
4	RN1, RN2, PhD, MD	All elements	0.63	0.77	0.84
4	RN1, RN2, PhD, MD	Outcome	0.78	0.94	0.97
5	PhD, MD	Outcome	0.75	—	0.75

Table 5 ■ Positive Predictive Value of the Stacking Meta-Classifier for Each Major Clinical Task

Task	Etiology	Diagnosis	Therapy	Prognosis
PPV	0.3	0.37	0.31	0.36

improvement (map = 0.4131) over the presentation order of the citations in PubMed retrieval results (map = 0.1425). We consider the number of relevant documents displayed at the top of a ranked list, measured as precision at rank 10, to be a more meaningful measure for point-of-service information delivery. Number of relevant documents in the first 10 of the PubMed retrieval results and after the re-ranking is shown in Table 6 along with the total number of the retrieved citations. A 389% improvement was achieved in precision at rank 10.

Discussion

In this study of the automatic outcome identification, we obtained encouraging results. In the intrinsic evaluation of the outcome classifiers, our automatic classification achieves results that approach human agreement (kappa value between the truth set and the system = 0.67 compared with kappa ranging from 0.8 to 0.94 in the upper bound established based on the outcome identification by clinicians). The systems also improve over the baseline at the three-sentence cutoff level. Similar to observations of Mendonça and Cimino,[23] the best performance is achieved for therapy and the worst for etiology. The stacking method that combines the base classifiers in an optimum way outperforms the one based on manually crafted rules in all cases but therapy. The success of the manual rules for therapy is understandable because the rules were created using 275 therapy-oriented citations that constitute the training set. The number of false positives in the highly ranked sentences may be reduced (at the cost of reducing sensitivity) by establishing high threshold, and marking as outcomes only sentences with high outcome confidence scores. The high outcomes confidence scores also serve as the foundation for the outcome-based ranking used in our extrinsic evaluation. As seen in Table 6, outcome-based ranking significantly outperforms PubMed Clinical Queries, presently one of the best tools available to clinicians. We are currently exploring application of the outcome identification system in answering clinical questions and organizing retrieved results within the framework of an EBM model. The identified outcome statements could be used to focus clinicians' attention when the devices they are using, for example, hand-held computers, are not capable of displaying the whole abstract in one screen. For a practical implementation, the outcome statements need to be identified in advance because processing of the retrieval results might require several minutes depending on the size of the result set.

Similar to findings of Rosenbloom et al.,[48] reliability studies conducted to verify validity of the test collection we created revealed that identification of outcome statements was straightforward for the RN and the MD with excellent inter- and intra-annotator agreement. Good agreement was also achieved among all experienced clinicians. The PhD annotator felt less confident reviewing biomedical literature and spent more time annotating the abstracts; however, agreement among all annotators was still good. The greatest disagreement in identification of the outcome statements occurred when the patient outcome was stated as a hypothesis or did not have supporting evidence in the abstract; for example, the following statement: "HIV-infected children older than 2 years would benefit from Hib vaccination although, one dose catch-up schedule is not sufficient in a third of these children" was annotated as an outcome by one of the RNs, and as a supposition by the other. Another source of disagreement was the length of the outcome statement (Fig. 1). Two categories that caused most disagreement were *intervention* and *supposition* (each with kappa = 0.4): the annotators had difficulties identifying preventive measures and epidemiologic techniques as interventions; and the annotators' background knowledge permitted recognizing suppositions not easily identified by the others (one of the RNs specialized in pediatric critical care, and the other has expertise in diabetes).

Should a future annotation effort be undertaken, the present study provides several observations that might improve interannotator agreement: (1) the schema should be revised, discussed, and understood by all annotators in advance, with special attention to the categories that caused most disagreement; (2) all annotators should be trained in critical appraisal of the medical literature and be familiar with the principles of EBM; and (3) all experts should annotate each document. In addition, the annotation schedule should provide for reconciliation of differences.

Our study has several limitations. First, although we tried to achieve independence from the topic of the study and the genre of the publication by diversifying our topics and search strategies and annotating all publication types, it is possible that the relatively small number of annotated documents and annotators introduced a selection bias and the system is not as generalizable as we hope. For example, the emerging topics such as pharmacogenetics were considered of no immediate clinical interest; however, this perception might change in the future, and the system will need to recognize as an outcome the following statement: "The presence of combined alleles M1 and T1 deficiencies in glutathione-*S*-transferase genes increases the susceptibility to tacrine hepatotoxicity."

Table 6 ■ Number of Relevant Documents in the First 10 of the PubMed Retrieval Results (PM), and after the EBM Model-Based Reranking (EBM)

	Back Pain		Obesity		Osteoporosis		Panic Disorder		Warts	
Ranking	PM	EBM	PM	EBM	PM	EBM	PM	EBM	PM	EBM
Relevant in 10	3	10	0	7	1	9	5	9	0	9
Precision at rank 10	0.3	1	0	0.7	0.1	0.9	0.5	0.9	0	0.9
Total retrieved	246		181		513		268		104	

Another limitation of the study is using PubMed Clinical Queries as the baseline system in the extrinsic evaluation. Although the retrieval results are among the best currently available to clinicians, one of the state-of-art search engines that rank retrieval results will be used as a baseline in our future experiments. Another technical limitation is introduced by the base semantic classifier: identification of semantic types in retrieval results takes time, which means the processing needs to be done off-line either in advance, for interactive information retrieval and question answering systems, or after the search in a summarization system that will provide digests of the searches asynchronously. At present, we are implementing a prototype EBM-based system that generates overviews of PubMed retrieval results. Once the system is implemented, it will be evaluated for its capability to facilitate clinical question answering similar to experiments presented in.[14] We plan to use outcome statements appraised by the users in these experiments for improvement of the outcome identification system.

Conclusions

This research presents a new EBM-based approach to processing and understanding MEDLINE citations used to meet clinicians' just-in-time information needs. The potential usefulness of a citation in a clinical setting is approximated using the presence of an outcome statement in the citation. We took a micro-level approach, i.e., finding specific outcome statements by evaluating each sentence in medical text, since our previous research using a macro-level approach, i.e., classifying a text as a whole to determine the presence of an outcome statement, showed only moderate associations between perceived clinical value of a citation and features that characterize the whole citation.[8]

Good agreement between annotators in identifying the outcome statements on the micro-level, i.e., annotating each sentence in the citation as belonging to one of the EBM model fields, motivated our in-depth study and implementation of the automatic methods for outcome identification. The development of an annotation scheme, a collection of MEDLINE citations annotated at the sentence level, and identification of essential textual, structural, and meta-information features for automatic classification of outcome sentences permitted the development of automatic outcome identification methods. The automated system combines domain knowledge and modern statistical methods to achieve performance in outcomes identification approaching that attained by human annotators. Our preliminary results in outcome-based document ranking show the potential validity of the EBM-based approach in delivering information critical to clinical decision support in a timely manner.

We would like to thank Malinda Peeples for her participation in the annotation and in the development of the annotation guidelines, and Sigmund Perez for annotation of the abstracts. We are very grateful to JAMIA reviewers for their detailed and thoughtful comments.

References ■

1. Arndt KA. Information excess in medicine. Overview, relevance to dermatology, and strategies for coping. Arch Dermatol. 1992; 128:1249–56.
2. Bakken S, Cimino JJ, Hripcsak G. Enabling patient safety and evidence-based practice through informatics. Med Care. 2004; 42:S49–56.
3. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ. 1996;312(7023):71–2.
4. Covell DG, Uman GC, Manning PR. Information needs in office practice: are they being met? Ann Intern Med. 1985;103:596–9.
5. Northup DE, Moore-West M, Skipper B, Teaf SR. Characteristics of clinical information–searching: investigation using critical incident technique. J. Med Educ. 1983;58:873–81.
6. Timpka T, Ekstron M, Bjurulf P. Information needs and information seeking behaviour in primary health care. Scand J Prim Health Care. 1989;7:105–9.
7. Williamson JW, German PS, Weiss R, Skinner EA, Bowes F III. Health science information management and continuing education of physicians. A survey of U.S. primary care practitioners and their opinion leaders. Ann Intern Med. 1989;110:151–60.
8. Demner-Fushman D, Hauser S, Thoma G. The role of title, metadata and abstract in identifying clinically relevant journal articles. Proceedings of the AMIA Fall Symposium. 2005;191–295.
9. Lewis DD. Evaluating and optimizing autonomous text classification systems. Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR. 1995:246–54.
10. Sebastiani F. Machine learning in automated text categorization. ACM Computing Surveys. 2002;1(34):1–47.
11. McKnight L, Srinivasan P. Categorization of sentence types in medical abstracts. Proceeding of the American Medical Informatics Association Annual Symposium. AMIA. 2003:440–4.
12. Cogdill Keith W. Information needs and information seeking in community medical education. Acad Med. 2000;75:484–6.
13. De Groote SL, Dorsch JL. Measuring use patterns of online journals and databases. J Med Library Assoc. 2003;91:231–41.
14. Hersh WR, Crabtree MK, Hickam DH, Sacherek L, Rose L, Friedman CP. Factors associated with successful answering of clinical questions using an information retrieval system. Bull Med Library Assoc. 2000;88:323–31.
15. Light M, Qiu XY, Srinivasan P. The language of bioscience: facts, speculations, and statements in between. Linking biological literature, ontologies, and databases. BioLINK. 2004;17–24.
16. Teufel S, Moens M. Summarizing scientific articles: experiments with relevance and rhetorical status. Comput Linguist. 2002;28: 409–45.
17. Hachey B, Grover C. A rhetorical status classifier for legal text summarization. In Proceedings of the ACL-2004 Text Summarization Branches Out Workshop. ACL-2004.
18. Bishop AP. Document structure and digital libraries: how researchers mobilize information in journal articles. Inform Process Manage. 1999;35:255–79.
19. Purcell G, Rennels G, Shortliffe E. Development and evaluation of a context-based document representation for searching the medical literature. Int J Digit Libraries. 1997;1:288–96.
20. Niu Y, Hirst G. Analysis of semantic classes in medical text for question answering. Proceedings of the ACL 2004 Workshop on Question Answering in Restricted Domains. ACL 2004.
21. Lin J, Quan D, Sinha V, et al. What makes a good answer? The role of context in question answering. Proceedings of the Ninth IFIP TC13 International Conference on Human-Computer Interaction. INTERACT 2003.
22. McKeown K, Chang SF, Cimino JJ, Feiner S, Friedman C, Gravano L, et al. PERSIVAL, a system for personalized search and summarization over multimedia healthcare information. Proceedings of the Joint Conference on Digital Libraries (JCDL). 2001;331–40.
23. Mendonça EA, Cimino JJ. Automated knowledge extraction from medline citations. Proceedings of the AMIA Fall Symposium. 2000:575–9.
24. Geissbuhler A, Miller RA. Clinical application of the UMLS in a computerized order entry and decision-support system. Proceedings of the AMIA Annual Symposium. 1998:320–4.

25. Larkey LS, Croft WB. Combining classifiers in text categorization. Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1996:289–97.
26. Lewis DD, Schapire RE, Callan JP, Papka R. Training algorithms for linear text classifiers. Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1996:298–306.
27. Ruiz ME, Srinivasan P. Combining machine learning and hierarchical indexing structures for text categorization. In Advances in Classification Research, Vol 10. Proceedings of the 10th ASIS SIG/CR Classification Research Workshop. ASIS 1999.
28. Dietterich TG. Ensemble methods in machine learning. Lecture notes in computer science. Cagliari (Italy): Springer, 2000.
29. Ting KM, Witten IH. Issues in stacked generalization. J Artif Intell Res. 1999;10:271–89.
30. Bennet P, Dumais S, Horvitz E. Probabilistic combination of text classifiers using reliability indicators: models and results. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR 2002.
31. Richardson W, Wilson MC, Nishikawa J, Hayward RSA. The well-built clinical question: a key to evidence-based decisions. ACP Journal Club. 1995;123:A-12.
32. Haynes RB, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. J Am Med Inform Assoc. 1994;1: 447–58.
33. Wilczynski NL, Haynes RB, Hedges Team. Robustness of empirical search strategies for clinical content in MEDLINE. Proc AMIA Symp. 2002;904–8.
34. Flaherty RJ. A simple method for evaluating the clinical literature. Fam Pract Manage. 2004;11(5):47–52.
35. WEKA: http://www.cs.waikato.ac.nz/~ml/weka/.
36. McCallum AK. "MALLET: a machine learning for language toolkit." http://mallet.cs.umass.edu. 2002.
37. Manning CD, Schütze H. Foundations of statistical natural language processing. Cambridge (MA): MIT Press, 1999.
38. Aronson AR. Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. Proc AMIA Symp. AMIA. 2001;17–21.
39. Yang Y, Pedersen JO. A comparative study on feature selection in text categorization. Proceedings of the 14th International Conference on Machine Learning. 1997:412–20.
40. Mladenic D. Feature subset selection in text learning. ECML'98. 1998;95–100.
41. McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Proceedings of 10th World Congress on Medical Informatics. MEDINFO. 2001.
42. USMLE Bulletin: http://www.usmle.org/bulletin/2005/2005bulletin.pdf.
43. Guyatt G, Rennie D (editors). Users' guides to the medical literature: essentials of evidence-based clinical practice. Chicago (Ill): AMA Press, 2002.
44. Sneiderman C, Demner-Fushman D, Fiszman M, Rindflesch TC. Semantic characteristics of MEDLINE citations useful for therapeutic decision-making. Proceedings of the AMIA Fall Symposium. 2005;1117.
45. Ebell MH, Siwek J, Weiss BD, Woolf SH, Susman J, Ewigman B, Bowman M. Strength of recommendation taxonomy (SORT): a patient-centered approach to grading evidence in the medical literature. J Am Board Fam Pract. 2004;17:59–67.
46. Harman D. Evaluation techniques and measures. In Proceedings of the 4th Text REtrieval Conference (TREC-4). 1996:A6–14.
47. Siegel S, Castellan N. Nonparametric statistics for the behavioral sciences. 2nd Ed. New York: McGraw-Hill, 1988.
48. Rosenbloom S, Giuse N, Jerome R, Blackford J. Providing evidence-based answers to complex clinical questions: evaluating the consistency of article selection. Acad Med. 2005;80:109–14.

APPENDIX 1

Search Strategies for Outcome Identification

Set 1

((((((("arthritis, rheumatoid"[MeSH Terms] OR RHEUMATOID ARTHRITIS[Text Word]) OR ("migraine"[MeSH Terms] OR MIGRAINE[Text Word])) OR ("breast neoplasms"[MeSH Terms] OR BREAST CANCER[Text Word])) AND (randomized controlled trial[Publication Type] OR ((randomized [Title/Abstract] AND controlled[Title/Abstract]) AND trial[Title/Abstract]))) AND jsubsetaim[text]) AND ("1999/1/1"[PDat] : "2004/1/1"[PDat]))

Set 2

("tuberculosis, pulmonary"[MeSH Terms] OR pulmonary tuberculosis[Text Word]) AND hasabstract[text] AND Randomized Controlled Trial[ptyp] AND English[Lang] AND ("human"[MeSH Terms] OR "hominidae"[MeSH Terms]) AND ("1999/01/01"[PDAT] : "2004/01/01"[PDAT])

("hypertension, renal"[MeSH Terms] OR renal hypertension[Text Word]) AND hasabstract[text] AND Randomized Controlled Trial[ptyp] AND English[Lang] AND ("human"[MeSH Terms] OR "hominidae"[MeSH Terms]) AND ("1999/01/01"[PDAT] : "2004/01/01"[PDAT])

("asthma, exercise-induced"[MeSH Terms] OR asthma, exercise-induced[Text Word]) AND hasabstract[text] AND Randomized Controlled Trial[ptyp] AND English[Lang] AND ("human"[MeSH Terms] OR "hominidae"[MeSH Terms]) AND ("1999/01/01"[PDAT] : "2004/01/01"[PDAT])

Set 3

(immunizations[Text Word] OR immunisations[Text Word] OR "immunization"[MeSH Terms]) AND hasabstract[text] AND English[Lang] AND ("infant, newborn"[MeSH Terms] OR "child, preschool"[MeSH Terms] OR "infant"[MeSH Terms]) AND ("adverse effects"[Subheading] OR adverse effects [Text Word])

Set 4

(diabetes mellitus[Text Word] OR "diabetes mellitus"[MeSH Terms] OR diabetes insipidus[Text Word] OR "diabetes insipidus"[MeSH Terms] OR diabetes[Text Word]) AND hasabstract[text] AND English[Lang] AND ("human" [MeSH Terms] OR "hominidae"[MeSH Terms])

Set 5

"treatment outcome"[MeSH Terms] AND "loattrfree full text"[sb] AND hasabstract[text] AND Randomized Controlled Trial[ptyp] AND English[Lang] AND "humans"[MeSH Terms]