NIH Public Access
Author Manuscript
J Biomed Inform. Author manuscript; available in PMC 2010 October 1.

Published in final edited form as:
J Biomed Inform. 2009 October ; 42(5): 801–813. doi:10.1016/j.jbi.2008.10.002.

Automatic Summarization of MEDLINE Citations for Evidence–Based Medical Treatment: A Topic-Oriented Evaluation

Marcelo Fiszman, M.D., Ph.D.[1], **Dina Demner-Fushman, M.D., Ph.D.**[1], **Halil Kilicoglu, M.S.**[1,2], and **Thomas C. Rindflesch, Ph.D.**[1]

[1] National Library of Medicine National Institutes of Health, Bethesda, MD

[2] Department of Computer Science and Software Engineering Concordia University, Montréal, QC

Abstract

As the number of electronic biomedical textual resources increases, it becomes harder for physicians to find useful answers at the point of care. Information retrieval applications provide access to databases; however, little research has been done on using automatic summarization to help navigate the documents returned by these systems. After presenting a semantic abstraction automatic summarization system for MEDLINE citations, we concentrate on evaluating its ability to identify useful drug interventions for fifty-three diseases. The evaluation methodology uses existing sources of evidence-based medicine as surrogates for a physician-annotated reference standard. Mean average precision (MAP) and a clinical usefulness score developed for this study were computed as performance metrics. The automatic summarization system significantly outperformed the baseline in both metrics. The MAP gain was 0.17 ($p < 0.01$) and the increase in the overall score of clinical usefulness was 0.39 ($p < 0.05$).

Keywords

Natural Language Processing; Semantic Processing; Automatic Summarization; Evidence-Based Medicine; Knowledge Representation; Artificial Intelligence; Evaluation

1. INTRODUCTION

The clinical research literature, particularly studies reporting on randomized clinical trials, provides an important information resource supporting effective patient care [1,2,3,4]. Compelling evidence that is most relevant to a particular disease is retrieved from online resources, especially MEDLINE, the National Library of Medicine's bibliographic database and the primary repository of the scientific literature. However, as such resources grow, it is increasingly challenging for clinicians to rapidly find useful answers to questions that arise during the course of practice.

Search engines and biomedical information retrieval techniques provide increased accuracy, ranking techniques, and ways of presenting results [5,6,7,8,9,10] to the biomedical researcher and clinician. However, little research has been published in using automatic summarization

Corresponding author: Marcelo Fiszman, M.D., Ph.D. National Library of Medicine 8600 Rockville Pike Bldg 38A, Rm, B1N-28J Bethesda, MD 20894 Phone: (301) 496 9305 Fax: (301) 480 3035 fiszmanm@mail.nih.gov..

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

to augment these techniques and help manage the information contained in the large numbers of MEDLINE citations often returned by PubMed searches. Automatic summarization seeks to provide the most important information from a source in a condensed format. This ability could support the practice of evidence-based medicine by allowing, for example, users to compare and contrast several treatments for a particular disease [11]. We are developing an automatic summarization system in the semantic abstraction paradigm [12] that can potentially help clinicians find the most salient information relevant to some disease. The thrust of the research reported here is to evaluate the summaries produced, in an effort to determine how useful they are in helping clinicians provide quality patient care.

We conducted a formal, large-scale, topic-based evaluation of our automatic summarization system, which found interventions in the biomedical literature for several questions about disorders. The questions and synthesized answers were semiautomatically extracted from the June 2004 issue of Clinical Evidence (CE) *concise*, a widely accepted resource for evidence-based medicine compiled by the British Medical Journal [13]. In addition, we enhanced this resource with disease-drug information from the Physicians' Desk Reference (PDR) [14], which provides access to FDA-approved interventions for over 4,000 drugs.

2. BACKGROUND

2.1 Semantic abstraction summarization system

In research on automatic summarization a contrast is made between processing a single text and several documents. To be usefully applied in managing the results of PubMed searches, multidocument applications are needed. Such systems have been discussed in general and in specific domains. Teufel [15] and Kupiek et, al [16] developed systems to summarize scientific articles in general. Several systems in the computational linguistics literature focus on current events [17,18,19] and other domain areas, such as legal documents [20]. An example in medicine is the PERSIVAL system [21]. Afantenos' [22] survey of summarization in the biomedical domain points out the popularity of the extraction paradigm [23], in which summaries consist of salient text identified in source documents. Other systems use semantic information to identify salient topics in scientific articles [24], to generate summaries for news articles [25], and to generate summaries of consumer health documents as well as technical articles for physicians [26]. Earlier work investigated the semantic abstraction paradigm [27], in which a summary is constructed from an abstract representation of the semantic content of source documents.

We are developing a summarizer in the semantic abstraction paradigm, and the system relies on semantic representation provided by SemRep [28,29], a natural language processing application under development at the National Library of Medicine. SemRep extracts semantic predications from the biomedical research literature based on two principles: underspecified linguistic analysis and domain knowledge from the Unified Medical Language System® (UMLS)® [30]. For example, SemRep interprets (1) as (2), where the arguments in this semantic predication, “Donepezil” and “Alzheimer's Disease,” are Metathesaurus concepts, and the predicate, TREATS, is from the Semantic Network.

(1) Donepezil for the management of Alzheimer's disease.

(2) Donepezil TREATS Alzheimer's Disease

The predications produced by SemRep comprise executable knowledge representing semantic information in the documents processed and can be reduced by the summarization process to provide an overview of those documents from four points of view (treatment of disease, diagnosis of disease, pharmacogenomics, and substance interactions). In this paper, we used the treatment of disease point of view.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Summarization relies on a user-specified topic and a transformation phase based on four principles which ensure that the summary provides useful information on the topic [12]. The principles are informally defined as:

(3) **Relevance**: Include predications on the topic of the summary **Connectivity**: Also include "useful" additional predications **Novelty**: Do not include predications that the user already knows **Saliency**: Only include the most frequently occurring predications

If "pneumonia" is selected as the topic of the summary, Relevance processing keeps predications with "Pneumonia" as an argument, such as "Ampicillin TREATS Pneumonia" but excludes all others (for example, "Doxorubicin TREATS Hodgkin's Disease"). Connectivity includes predications that share an argument with a predication kept by Relevance (such as "Ampicillin CAUSES Rash"). Novelty uses the hierarchical structure of the Metathesaurus to eliminate predications with generic (and hence uninformative) arguments such as "Pharmaceutical Preparation TREATS Pneumonia." Finally, Saliency eliminates predications with low frequency of occurrence [31].

A high level view of the summarization process evaluated in this study is illustrated in Figure 1. The results of a PubMed search are first interpreted by SemRep. The summarizer then takes these predications as input along with a user-specified topic (such as a disease) and applies the transformation principles to produce a reduced set of predications (or conceptual condensate), which gives an overview of the content of the retrieved citations. Finally, the condensate is represented as a graph, which is both informative, in the sense that it provides an overview of the content of the source citations, and indicative, in that predications are linked to the source text that generated them.

An example of output from this system is given in Figure 2, where the graph represents semantic predications produced by summarizing 300 MEDLINE citations returned by a PubMed search on panic disorder. Nodes in the graph represent concepts, and arcs show relations between them. (Only TREATS relations are displayed in Figure 2.) Drug therapies, such as selective serotonin reuptake inhibitors, tricyclic antidepressants (imipramine), and benzodiazepines appear in the graph and are included as effective interventions in CE. Perhaps more interestingly, cognitive therapy and psychotherapy, of particular interest for anxiety disorders, are prevalent in the summary but are not included in the June 2004 release of CE. (In that version it is noted that these two interventions will be considered for future releases.) In Figure 2, the arc linking "Cognitive Therapy" to "Panic Disorder" has been selected and the information in the box on the right indicates that this TREATS relation occurs 17 times (*frequency of occurrence*) in 11 of the citations returned by the search (*typicality*). One of these citations is shown in the box at the bottom; it reports on a multicenter trial of different delivery methods of cognitive-behavioral therapy for the treatment of panic disorder.

2.2 Related research on evaluating multidocument summarization

Evaluation of multidocument summarization is an evolving research field. In most studies [32,33,34], reference standard summaries are produced by several experts, and measures of intra-and inter-rater agreement are provided. The summarizers are contrasted quantitatively with these reference standards and several performance measures are computed. This methodology is expensive and time consuming. In addition, generating an ideal summary is subjective with respect to the experts who produce them. Because of this, research is being pursued to produce reference standards automatically [35].

Other evaluation studies are user centered. Typically this method has been deployed for single document summarization [36] and seeks to assess a summary based on how well a user can exploit it to perform a given task. Evaluation studies of multidocument summarization in the biomedical [37] and news (for example [38]) domains attempt to measure the impact that

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

multidocument summarization has on a user's ability to find answers quickly while satisfying information needs. After reading summaries, users are asked to answer questions related to finding required information, exploiting that information, and user satisfaction.

More recently Amigo et al. [39] proposed an “information synthesis” task, defined as “given a specific information need, the multidocument summary should extract, organize, and synthesize an answer that satisfies that need.” Based on this proposal, the annual Document Understanding Conference (DUC) [40] was reengineered to address a more focused, topic-oriented approach to evaluating automatic summarization systems. The topic invokes real-world questions and human assembly of answers so they can be compared against the results of the summarizers.

In this study, we followed the information synthesis approach of Amigo et al. in an effort to determine how interventions that appear in our summaries as subject arguments of TREATS predications compare with interventions in a reference standard generated semiautomatically from two widely accepted resources for evidence-based medicine, CE and PDR. Therefore, we took advantage of these resources as surrogates for answers assembled by hand, and very little human involvement was necessary in the creation of the reference standard. These resources will be briefly introduced in the next section, and compilation of the reference standard will be discussed in detail in the Methods section.

2.3 Resources for evidence based medicine

The June 2004 issue of CE was used to create the reference standard for evaluation of the generated summaries. CE is one of the so-called secondary sources of evidence-based information for clinicians; such sources provide concise regularly updated summaries of the best available clinical research. CE summaries are presented as answers to clinical questions. Each summary includes a list of interventions, key points and synopses of the reviewed clinical studies, references, and supplements for a particular disorder. The usefulness of this resource comes from the ordered categorical output of the list of interventions. For example, for the question “What are the treatments for osteoarthritis?” CE has an ordered list of interventions such as oral analgesics, oral nonsteroidal antiinflammatory agents, exercise, etc. This structure of CE facilitates semiautomatic extraction of interventions needed to compile a reference standard. The ordered categories defined for CE interventions are:

Beneficial (Interventions for which effectiveness has been demonstrated by clear evidence from randomized clinical trials and for which expectation of harm is small compared with the benefits)

Likely to be beneficial (Interventions for which effectiveness is less well established than for those listed under “beneficial”)

Trade-off between benefits and harms (Interventions for which clinicians and patients should weigh up the beneficial and harmful effects according to individual circumstances and priorities)

Unknown effectiveness (Interventions for which there are currently insufficient data or data of inadequate quality)

Unlikely to be beneficial (Interventions for which lack of effectiveness is less well established than for those listed under “likely to be ineffective or harmful”)

Likely to be ineffective or harmful (Interventions for which ineffectiveness or harmfulness has been demonstrated by clear evidence)

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

In addition to CE, we used the 2004 version of the PDR, which lists drugs approved by the FDA for the treatment of disease. It does not provide ordered categories, but it compensates for the intentional sparseness of CE in a manner that will be explained in the next sections.

3. METHODS

For our topic-based approach to automatic summarization evaluation, we adapted methods developed by the National Institute of Standards and Technology for the DUC and Text Retrieval Conference (TREC). These evaluations rely on: 1) topics (descriptions of complex information needs, recently expressed as questions), 2) documents for summarization, 3) reference standards, and 4) evaluation metrics. We used topics and summaries compiled by medical experts and published in CE. In our collection, each topic is a question about pharmacotherapy for a given disorder. The corresponding summary answers this question by providing a ranked list of therapeutic interventions. We evaluated the system using widely accepted measures of performance developed in TREC and an evaluation metric developed for this study. The latter metric strives to capture the usefulness of the summaries for evidence-based medicine. The results from our summarization system were compared to a baseline in which answers were based solely on frequently occurring drugs in retrieved documents. Finally, we conducted a manual evaluation of four randomly selected topics.

3.1 Topics and reference standard

All questions from the CE 2004 issue pertaining to pharmacological treatment of diseases were included in this study (questions about therapeutic procedures such as surgery were excluded). Of the 192 CE topics, fifty-four questions matched the inclusion criterion. The corresponding topics (disorders) are listed in Table 1 along with the disease classes to which they belong (determined by CE). The fifty-four topics are expressed as follows (with minor variations) in CE: What are the effective treatments for X? (X represents disorders such as chronic obstructive pulmonary disease.)

The reference standard intervention lists for the fifty-four topics were formed as follows: The core of the reference standard consists of the interventions extracted manually from the CE 2004 issue. CE provides information for an international audience; therefore, only the generic or the most common names of drugs (rather than brand names) are used in the summaries. Further, CE does not include all FDA-approved drugs. To compensate for this sparseness, drugs for each topic were manually extracted from PDR and added to the reference standard as a separate group. Including PDR drugs also addressed the fact that CE does not include brand names of drugs. Each CE drug name in the reference standard was annotated with its category (such as “beneficial,” “trade-off,” or “likely harmful”), while the PDR drugs were labeled “pdr.” For a given topic, if both sources (CE and PDR) contained a drug name, only its CE category was used.

After compiling lists of drugs for each topic, the reference standard was normalized by mapping drug names to UMLS Metathesaurus concepts using MetaMap [41]. To provide for subsequent automatic evaluation (in which drugs found by the summarization system and the baseline method were matched against the reference standard on a conceptual level), drug names were represented in the reference standard using UMLS unique concept identifiers (CUI). The majority of the UMLS CUIs for the reference standard were determined automatically; however, CUIs for a few unmatched drugs were assigned by the second author (DDF), who was not involved in the development of the summarization system. Manual assignment of CUIs was necessary to accommodate spelling variation and synonymy not represented in the Metathesaurus and to disambiguate multiple matches. The reference standard was created independently and in advance of the summaries generated by the summarization system and the baseline method. Table 2 presents the reference standard entry for the topic panic disorder.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

3.2 Documents for summarization

MEDLINE citations for the summarization system and the baseline were retrieved automatically by submitting a query template (Figure 3) to PubMed using EUtilities [42]. The template emulates search strategies employed by medical librarians seeking high quality information focused on a particular disease: 1) To focus on the disease in question, the default PubMed query expansion was turned off using the [mh:noexp] command. For example, this strategy prevents broadening the search for “epilepsy” to a search for 25 disorders, including “Seizures,” “Febrile Seizures,” and “Landau-Kleffner Syndrome.” 2) To focus on high quality evidence, the search was limited to the results of clinical trials. Although meta-analyses and systematic reviews are considered high quality secondary sources of evidence, we followed the strategies developed by the authors of these studies and focused on reliable primary sources of evidence. Further, to be as close as possible to knowledge available at the time of the creation of CE, we restricted the date of the search to articles published prior to the publication of the CE article for a given disease. To fill the term variable in this template, each topic in Table 1 was mapped to MeSH using MetaMap. We found that a few of the resulting MeSH concepts did not adequately convey the meaning of the topic, and we therefore modified the names of these topics and remapped them to MeSH. For example, initially “oropharyngeal candidiasis” was mapped to MeSH “Candidiasis,” while the manually modified “oral candidiasis” was mapped to the more specific “Candidiasis, Oral.” The modifications are as follows:

oropharyngeal candidiasis → oral candidiasis

chronic bacterial prostatitis

→ chronic prostatitis

chronic plaque psoriasis → chronic psoriasis

early stage aggressive non-Hodgkin's lymphoma → early stage non-Hodgkin's lymphoma

The PubMed query based on the template in Figure 3 is general and was used for the 54 topics of our evaluation. The number of citations returned varied depending on the topic. For example, there were no results of clinical trials published specifically on “Irritable Bowel Syndrome” before 2004. Since the search returned no citations, this topic was ignored for subsequent processing, and the study was conducted with fifty-three topics. There were few trials published on four other topics (early stage aggressive non-Hodgkin's lymphoma, congenital toxoplasmosis, generalized anxiety disorder, and leg cramps), and search results were consequently low. Although results for these topics were included in the study, summarization system results were affected.

3.3 Baseline generation

To evaluate the usefulness of the summarization system for evidence-based practice, we compared it to a baseline in which summaries were generated using simple frequency of occurrence of drug names in retrieval results. The baseline was created using MetaMap. Concepts having a semantic type in the UMLS semantic group Chemicals & Drugs were identified in the PubMed search results for each of the fifty-three topics in our study (excluding “Irritable Bowel Syndrome”). The five most frequently occurring drugs in the set of documents for each disorder were then extracted as the list of drugs considered to be pharmacotherapies for that disorder. The five most frequently occurring drugs were selected, because, on average, there are five *beneficial* and *likely beneficial* drugs listed in CE for each disorder. The baseline method, emphasizing frequency of occurrence, may approximate a busy clinician investigating therapeutic alternatives when confronted with a particular disease.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

3.4 Summaries generated by the summarization system

For each of the fifty-three topics in the study, the retrieved document set was processed using SemRep, and the predications returned were summarized with the relevant topic specified as the main topic of the summary. From the summarized conceptual condensate we extracted predications of the form <Intervention> TREATS <Disorder> where <Disorder> is the UMLS Metathesaurus concept corresponding to the topic, and <Intervention> is any Metathesaurus concept having a semantic type in the semantic group Chemicals & Drugs. Final ranked lists of drug therapies for each topic were created by extracting <Intervention> concepts and sorting them by frequency of occurrence.

For four topics (early stage aggressive non-Hodgkin's lymphoma, congenital toxoplasmosis, generalized anxiety disorder, and leg cramps) the summarization system produced no results because the number of citations retrieved for these topics was small (as noted above), leaving the output for summarization either empty or with no TREATS relations. For example, PubMed retrieved two citations for congenital toxoplasmosis; there were no TREATS predications in the summarization output for these citations.

3.5 Evaluation

Determining whether a drug name found by the summarization system belongs to the reference standard as a treatment for the relevant topic is fairly complex because of the nature of CE, on which the reference standard was based. Different, but synonymous, drug names occurring in the reference standard and retrieved by the summarization system are accommodated by synonymous concepts in the UMLS Metathesaurus. However, in many cases the reference standard names a class of drugs as beneficial for a given topic, for example, thrombolytic agents for acute myocardial infarction. In this case, any thrombolytic agent found by the summarization system should be counted as a true positive. The hierarchical matching algorithm described in section 3.5.1 performs this task automatically. Hierarchical matching is followed by computation of mean average precision (section 3.5.2). Although this metric is useful in predicting future performance of a system with respect to finding drugs mentioned in the reference standard, it does not rate beneficial drugs higher than harmful ones. We therefore developed a metric, the clinical usefulness score, which takes into consideration the quality of the intervention found (section 3.5.3). We computed mean average precision and the clinical usefulness score for each disease topic and then averaged the scores within each disease class. An overall schematic view of our evaluation is depicted in Figure 4.

3.5.1 Hierarchical matching—As the basis for hierarchical matching, we used the UMLS Knowledge Source Server [43] to retrieve Metathesaurus hierarchical contexts for drugs returned by the summarization system and the baseline method. For example, “Thrombolytic Agents” was computed as an ancestor of “Tissue Plasminogen Activator.” Drug names returned by the summarization system or the baseline method were allowed to match their ancestors, thus reconciling a drug name returned by the summarization system with its class in the reference standard. However, a class from the summarization system (acetylcholinesterase inhibitors, for example) compared to a member in the reference standard (donepezil) was marked as a false positive.

3.5.2 Mean average precision—Mean average precision (MAP) is a measure sensitive to the ranking of drugs by a system and summarizes both recall and precision [44]. MAP for the fifty-three topics was computed as the mean of the individual average precision scores for each topic. Average precision of each topic is the mean of the precision scores computed after each reference standard drug is found in the ranked list of drugs generated by the summarization system and the baseline method. Based on the fifty-three disease topics, we calculated MAP

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

for the eighteen disease classes presented in Table 1 by computing the average MAP scores for the topics in each class.

3.5.3 Clinical usefulness score—To evaluate the usefulness of the summarization system and the baseline method in a clinical setting, a categorical performance metric was developed specifically for this study. In calculating this score, interventions extracted by a system are assigned to one of four high-level categories depending on how they match the interventions in the reference standard. The goal is to give credit to the system for finding beneficial interventions and, similarly, penalize it for finding harmful interventions. The high-level categories and the corresponding reference standard categories are as follows:

BEST: beneficial, likely to be beneficial

OK: trade-off between benefits and harms, pdr

BAD: likely to be ineffective or harmful

OTHER: unknown effectiveness, unlikely to be beneficial

We compared summarization and baseline results with the reference standard using hierarchical matching and assigned each intervention extracted to one of these high-level categories. We then computed scores for each of these four categories for each disease class of Table 1. The score is normalized by dividing the number of interventions in the category by the total number of interventions extracted by the summarization process. An overall score of clinical usefulness is computed as follows:

$$\text{Overall score} = w_b^* \text{ BEST score} + w_{ok} \text{ OK score} - w_o \text{ OTHER score} - w_p^* \text{ BAD score}$$

Where:

D: total # of drugs from system (summarization or baseline)

BEST score = (# of beneficial + # of likely beneficial) / D

OK score = (# of trade-off + # of pdr) / D

BAD score = (# of likely harmful) /D

OTHER score = 1 − (BEST score + OK score + BAD score)

The BEST category is given more weight (w_b coefficient) because it is more important to find a highly beneficial treatment than to find an intervention of questionable effectiveness (OK category). Similarly, the summarization system and the baseline method are penalized more (w_p coefficient for the BAD category) for finding a harmful intervention than for finding a treatment of unknown effectiveness (OTHER category). The overall score is a number that takes into consideration the degree of usefulness of an intervention. This score is potentially meaningful in evaluating relative performance of several systems on a given test set. As the weights w_b and w_p are constant, their absolute values represent an evaluator's belief in the importance of a given category. In this study both coefficients were (intuitively) set to 3, with the consequence that a drug with the BEST score is considered to be three times more useful than one with an OK score, which is in turn three times more useful than a drug with a BAD score.

We tested for statistically significant differences between the performance measures (mean average precision and clinical usefulness score) for the summarization system and the baseline for all disease classes using a two-tailed Wilcoxon signed rank test with 5% significance level.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

In addition, we used Kendall's tau [45] to determine whether mean average precision and the clinical usefulness score were correlated with each other.

3.5.4 Manual evaluation—Automatic comparison of system results to the reference standard is less resource intensive than relying on humans, but is not as accurate. In an effort to provide additional insight into the effectiveness of the summarization system for providing clinically relevant information, we conducted a manual evaluation for a random sample of four topics: acute myocardial infarction, gastroesophageal reflux disease, community-acquired pneumonia, and panic disorder. Intervention concepts returned by the summarization system were marked as being correct only if they matched a drug in the reference standard categories "beneficial" or "likely to be beneficial." Recall and precision were calculated by hand.

4. RESULTS

Table 3 (Appendix A) shows all the interventions found by the summarization system for each of the topics in the study. As mentioned before, irritable bowel syndrome was excluded, leaving the digestive system class with one topic, for a total of 53 topics in all classes. Also, as explained before, the summarization system produced no results for four topics (early stage aggressive non-Hodgkin lymphoma, congenital toxoplasmosis, generalized anxiety disorder, and leg cramps). For the fifty-three topics, the number of citations represented in the final summary varied from 2 to 500 with an average of 240. The number of interventions for the topics varied from 0 to 26 with an average of 9. In the last column of Table 3, the interventions retrieved for each topic are displayed in descending order of frequency of occurrence and typicality. All the interventions in Table 3 are UMLS Metathesaurus concepts.

Table 4 compares results from the summarization system to the baseline (determined exclusively by frequency of occurrence) and lists five interventions from each method that were found to treat the dementia topic (mental health class). Underlined are interventions considered beneficial ("Donezepil") and likely to be beneficial ("Gingko biloba extract") according to the reference standard and found by the summarization system, but not by the baseline method. The uninformative concept "Pharmaceutical Preparations" is eliminated by the summarization system but occurs in the baseline. The other interventions are "Antipsychotic agents" such as "Risperidone," "Haloperidol," and "Olanzapine." These are used to control behavioral and psychological symptoms of dementia, but do not improve the disease and are listed as "trade-off between benefits and harms" by the reference standard.

As noted, in order to assess effectiveness of the summarization system we calculated both MAP and a clinical usefulness score. Approaches to evaluating the system should reflect the intended task. For example, if the goal is to display a ranked list of treatments classified from useful to harmful, MAP better predicts future performance. However, if the task is to display only the top-ranked effective treatments, the clinical usefulness score is a better evaluation metric.

Table 5 shows mean average precision for the baseline method and summarization system on identifying reference standard interventions for the disease classes noted in Table 1, sorted in descending order of MAP gain between results from the summarization system and the baseline method (last column). MAP gain computed over all disease classes is statistically significant ($p < 0.01$).

Table 6 shows the results for the overall score of clinical usefulness for the baseline method and the summarization system. The negative values in the third and fourth columns indicate that the summarization system either completely failed to find the best available treatments, or the proportion of the beneficial drugs was insignificant compared to the number of harmful drugs found by the system. The last column of the table shows the gain in the overall clinical

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

usefulness score with respect to the baseline in descending order. The overall difference in usefulness score between results for the baseline method and the summarization system is statistically significant ($p < 0.05$). Although, both performance measures showed significant improvement over the baseline they only moderately correlate with each other (Kendall's tau = 0.34), which means they are measuring performance in different ways.

Tables 5 and 6 demonstrate that according to both scores the summarization system performed better than the baseline method for such disease classes as oral health, respiratory disorders, musculoskeletal disorders, mental health, and digestive system disorders. On the other hand, the summarization system produced degraded results with respect to the baseline for the HIV and AIDS disease class. Results for other disease classes vary depending on the performance measure used. In the next section we consider possible reasons for these results. We also discuss drugs found by the summarization system that occur in the literature as valid interventions for specified disorders, but which are not included in the reference standard.

The results of the manual evaluation performed are presented in Table 7. Recall is strong, but precision less so.

5. DISCUSSION

Overall, our results support the hypothesis that automatic summarization can make a positive contribution to effective clinical care by managing the information contained in MEDLINE citations relevant to specific topics. The overall statistically significant MAP gain with respect to the baseline was 0.17. The overall improvement in the clinical usefulness score was 0.39, which reflects the summarization system's ability to find interventions that have been proven beneficial or are likely to be beneficial according to the reference standard. These gains were achieved despite a relatively strong baseline, generated by using MetaMap conceptual normalization to identify drugs in citations retrieved with a focused search on specific topics.

The improvement in several disease classes (for example musculoskeletal disorders, mental health, and respiratory disorders) can be attributed to predication-based summarization. This method takes advantage of predications of the form “<Intervention> TREATS <Disorder>” produced by SemRep in order to focus drug therapies on the topic (Relevance processing). For example, if the question was for treatment of acute asthma, the summary will not retrieve interventions for chronic asthma. Further, in the Novelty phase, the summarization system prunes predications with arguments that are uninformative. Therefore, in general, a nonspecific concept such as “Pharmaceutical Preparations” (which does not appear in the reference standard) is always eliminated.

Although, the summarization system performed better than the baseline method for most disease classes, for others, results were unchanged or degraded. In some cases, degraded results were ultimately due to UMLS Metathesaurus coverage. For example, in the HIV and AIDS class (which has one disease, pneumocystis carinii pneumonia) both MAP and the clinical usefulness score were better for the baseline than for results from the summarization system. Both the summarization system and baseline processing retrieved four drugs that appear in the reference standard as beneficial in the treatment of pneumocystis carinii pneumonia in HIV patients: trimethoprim-sulfamethoxazole combination, pentamidine, corticosteroids, and dapsone. Additional interventions found by both the baseline method and the summarization system are not in the reference standard. Zidovudine, a drug used against the HIV virus, but not for the treatment of pneumocystis carinii pneumonia, appears in the baseline as a false positive. “Prophylactic treatment” and hydroxynaphthoquinone 566C80 were returned by the summarization system, neither of which is in the reference standard. Because the summarization system found one more false positive than the baseline method, it received a

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

lower performance measure than the baseline. In fact, hydroxynaphthoquinone 566C80 is a synonym for atovaquone, which appears in the reference standard as being beneficial for pneumocystis carinii pneumonia. Unfortunately, this equivalence is not represented in the Metathesaurus. Errors due to unnamed synonymy in the Metathesaurus degraded summarization system results in several other disease classes as well.

Mapping to the Metathesaurus during the process of generating semantic predications produced several errors. For example, the noun *dose* wrongly matched the Metathesaurus concept "DOS," which is a synonym for docusate, a stool softener. Similarly infelicitous mappings produced other false positives, such as "The science and art of healing" and "Stimulation – action."

The etiology of another class of errors involves curation policy in creating a secondary evidence-based resource such as CE. Interventions are included only when there is sufficient evidence to support a determination of their effectiveness. For example, although baclofen is discussed in the research literature as a promising drug for gastroesophageal reflux disease, it does not appear in the version of CE we used for the reference standard in this study (nor does it appear in PDR for this disease). Since our summarization system does not have access to CE curation policy, it retrieved baclofen as a treatment for gastroesophageal reflux disease based on a predication accurately extracted from the following text in MEDLINE: "Effect of acute and chronic administration of the GABA B agonist baclofen….in control subjects and in patients with gastrooesophageal reflux disease" (PMID 12631652).

In analyzing the results of the manual evaluation, false positives can be classified into four types. In descending order of frequency they are: intervention concepts that are categorized "trade-off or unknown effectiveness" (43%) or do not appear in the reference standard (29%), intervention concepts that are too general (21%), and infelicitous mappings to the UMLS Metathesaurus (7%). The majority of the errors were of the first two types and are related to CE curation policy. For example, "Antacids" (for gastroesophageal reflux disease) and "Alprazolam" (panic disorder) were retrieved by the summarization system but are classified in the "trade-off" or "unknown effectiveness" category. The second error type included interventions that do not appear in the reference standard, such as "Baclofen" (gastroesophageal reflux disease). These interventions are discussed in MEDLINE and may be included in later versions of CE. Errors of the other two types reflect shortcomings of the summarization system; however, they are less frequent than those in the first two types. "Antibiotics" (for community acquired pneumonia) is an example of a concept that is too general, and "Administration (procedure)" (acute myocardial infarction) is due to an incorrect mapping to the Metathesaurus.

The evaluation discussed here has several limitations. We only considered pharmacologic treatments, and did not address topics for which the primary intervention is a therapeutic procedure. In addition, CE does not include information on diagnosis or prognosis, and so the evaluation does not provide insight on system performance in these areas. Other curated resources (such as UpToDate) could be explored as a basis for extending the methodology discussed here beyond treatment. A further limitation is due not so much to the evaluation methodology as to the natural language processing system assessed, which does not have access to information about the quality of the evidence supporting the relevant intervention [46]. This kind of information supports CE intervention categorization, and taking it into account would likely decrease the number of false positives returned by the system (especially in the "trade-off" and "unknown effectiveness" categories). Research is being pursued on automatic determination of quality of evidence [47,48]; such processing could be incorporated into the summarization system.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Although this paper concentrated on evaluating automatic summarization with a view toward assisting clinicians in navigating MEDLINE, it is also attractive to consider the system described as a possible tool for developing secondary sources of information such as CE. The construction (and updating) of these resources is labor-intensive and expensive. The system presented above could potentially support curators in their work.

6. CONCLUSION

Physicians have access to an ever increasing number of online resources to support high quality patient care. Current research in biomedical information management technology provides several retrieval techniques to help find the most useful documents relevant to questions that arise during clinical practice. However, few studies have investigated automatic summarization as a potential tool to help navigate the retrieved documents. We describe a system based on semantic abstraction that summarizes MEDLINE citations discussing treatment for specified disorders. A graphical display gives an informative overview of the processed information, while links to the underlying documents allow access to details. This paper then concentrates on a formal evaluation of the accuracy of this automatic summarization system in identifying treatments for disorders.

We used a topic-oriented evaluation that follows the principle of “information synthesis” used in recent document understanding conference evaluations. As a surrogate for a physician-annotated reference standard, we semiautomatically compiled drug therapies for fifty-three topics from the June 2004 issue of Clinical Evidence *concise*, published by the British Medical Journal. This resource was enhanced with topic-drug information from the Physicians' Desk Reference. PubMed searches were issued for the fifty-three disorders studied, and the MEDLINE citations retrieved were processed by the summarization system. A baseline was also created by identifying the five most frequently occurring drugs in the citations retrieved for each disorder. Results from the summarization system and the baseline method were automatically compared to the reference standard, and two performance metrics were calculated: mean average precision and a clinical usefulness score, which penalized results that included drugs known to be harmful. The quality of automatic evaluation was checked through a manual assessment of the summarization results for four diseases. The summarization system scored significantly higher than the baseline on both metrics.

Acknowledgments

We would like to thank Charles Sneiderman and Jimmy Lin for valuable discussions about the evaluation methodology and available resources. This research was supported in part by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

REFERENCES

1. Wennberg JE. Unwarranted variations in healthcare delivery: Implications for academic medical centers. Br Med J 2002;325(7370):961–4. [PubMed: 12399352]
2. McGlynn EA, Asch SM, Adams J, et al. The quality of health care delivered to adults in the United States. N Engl J Med 2003;348(26):2635–45. [PubMed: 12826639]
3. Institute of Medicine, Committee on Quality Health Care in America. Crossing the quality chasm: A new health system for the 21st century. National Academy Press; Washington, DC: 2001.
4. Sackett, DL.; Straus, SE.; Richardson, WS.; Rosenberg, W.; Haynes, RB. Evidence-based medicine: how to practice and teach EBM. Churchill Livingstone; Philadelphia, PA: 2000.
5. Hersh W, Hickam DH, Haynes RB, et al. Evaluation of SAPHIRE: an automated approach to indexing and retrieving medical literature. Proc Annu Symp Comput Appl Med Care 1991:808–12. [PubMed: 1807718]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

6. Srinivasan P. Retrieval feedback in MEDLINE. J Am Med Inform Assoc 1996;3(2):157–67. [PubMed: 8653452]
7. Schardt C, Adams MB, Owens T, et al. Utilization of the PICO framework to improve searching PubMed for clinical questions. BMC Med Inform Decis Mak 2007;7(16)
8. Ide NC, Loane RF, Demner-Fushman D. Essie: a concept-based search engine for structured biomedical text. J Am Med Inform Assoc 2007;14(3):253–63. [PubMed: 17329729]
9. Demner-Fushman D, Lin J. Answering clinical questions with knowledge-based and statistical techniques. Computational Linguistics 2007;33(1):63–103.
10. Lin Y, Li W, Chen K, et al. A document clustering and ranking system for exploring MEDLINE citations. J Am Med Inform Assoc 2007;14(5):651–61. [PubMed: 17600104]
11. Mani, I. Automatic summarization. John Benjamins Publishing Co; Philadelphia, PA: 2001.
12. Fiszman M, Rindflesch TC, Kilicoglu H. Abstraction summarization for managing the biomedical research literature. Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics 2004:76–83.
13. Clinical evidence concise. BMJ Publishing Group; London: 2004.
14. Physicians' Desk Reference. Medical Economics Data; Montvale, NJ: 2004.
15. Teufel S, Moens M. Summarizing scientific articles - experiments with relevance and rhetorical status. Computational Linguistics 2002;28(4):409–445.
16. Kupiec, J.; Pedersen, J.; Chen, F. A trainable document summarizer; Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval; 1995. p. 68-73.
17. McKeown KR, Barzilay R, Evans D, et al. Tracking and summarizing news on a daily basis with Columbia's Newsblaster. Proc. of HLT- NAACL 2002:280–285.
18. Radev D. A Common theory of information fusion from multiple text sources, step one: cross-document structure. Proceedings of 1st ACL SIGDIAL Workshop on Discourse and Dialogue 2000:74–83.
19. Daniel N, Radev Dr, Allison T. Sub-event based multi-document summarization. Proceedings of the HLT-NAACL Workshop on Text Summarization 2003:9–16.
20. Grover C, Hachey B, Korycinski C. Summarising legal texts: sentential tense and argumentative roles. Proceedings of the HLT-NAACL Workshop on Text Summarization 2003:33–40.
21. McKeown HR, Chang SF, Cimino J, et al. PERSIVAL, a system for personalized search and summarization over multimedia healthcare information. Proceedings of the 1st ACM/IEEE-CS joint conference on digital libraries 2001:331–340.
22. Afantenos S, Karkaletsis V, Stamatopoulos P. Summarization from medical documents: a survey. Artif Intell Med 2005;33(2):157–77. [PubMed: 15811783]
23. Lin, CY.; Hovy, E. From single to multi-document summarization: a prototype system and its evaluation; Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics; 2001. p. 457-464.
24. Paice, CD.; Jones, PA. The identification of important concepts in highly structured technical papers; Proceedings of the 16th annual international ACM SIGIR conference on research and development in information retrieval; 1993. p. 69-78.
25. McKeown, K.; Radev, D. Generating summaries of multiple news articles; Proceedings of the 18th Annual International ACM SIGIR conference on research and Development in Information Retrieval; 1995. p. 74-82.
26. Elhadad N, Kan MY, Klavans JL, McKeown KR. Customization in a unified framework for summarizing medical literature. Artif Intell Med 2005;33(2):179–98. [PubMed: 15811784]
27. Hahn U, Mani I. The challenges of automatic summarization. Computer 2000;33(11):29–36.
28. Rindflesch, TC.; Fiszman, M.; Libbus, B. Semantic interpretation for the biomedical research literature. In: Chen, H.; Fuller, S.; Hersh, W.; Friedman, C., editors. Medical informatics: knowledge management and data mining in biomedicine. Springer; New York: 2005. p. 399-422.
29. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text. J Biomed Inform 2003;36(6):462–77. [PubMed: 14759819]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

30. Humphreys BL, Lindberg DA, Schoolman HM, Barnett GO. The Unified Medical Language System: an informatics research collaboration. J Am Med Inform Assoc 1998;5(1):1–11. [PubMed: 9452981]
31. Hahn, U.; Reimer, U. Knowledge-based text summarization: salience and generalization operators for knowledge base abstraction. In: Mani, I.; Maybury, MT., editors. Advances in automatic text summarization. MIT Press; Cambridge: 1999. p. 215-232.
32. Jing H, Barzilay R, McKeown KR, et al. Summarization evaluation methods: Experiments and analysis. AAAI Symposium on Intelligent Summarization 1998:60–68.
33. Lin CY, Hovy E. Manual and automatic evaluation of summaries. Proceedings of the ACL Workshop on Automatic Summarization 2002:45–51.
34. Radev, D.; Teufel, S.; Saigon, H., et al. Evaluation challenges in large-scale document summarization; Proceedings of the 41st Annual Meeting on Association for Computational Linguistics; 2003. p. 375-382.
35. Nenkova A, Passonneau R, McKeown K. The Pyramid Method: Incorporating human content selection variation in summarization evaluation. ACM Transactions on Speech and Language Processing 2007;4(2)
36. Mani, I.; House, D.; Klein, G. The TIPSTER SUMMAC text summarization evaluation; Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics; 1999. p. 77-55.
37. Elhadad N, McKeown K, Kaufman D, Jordan D. Facilitating physicians' access to information via tailored text summarization. AMIA Annu Symp Proc 2005:226–30. [PubMed: 16779035]
38. McKeown, H.; Passonneau, RJ.; Elson, DK., et al. Do summaries help? A task-based evaluation of multidocument summarization; Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval; 2005. p. 210-217.
39. Amigo, E.; Gonzalo, J.; Peinado, V. An empirical study of information synthesis tasks; Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics; 2004. p. 207-214.
40. http://duc.nist.gov/duc2005/tasks.html
41. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001:17–21. [PubMed: 11825149]
42. http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html
43. Bangalore A, Thorn KE, Tilley C, Peters L. The UMLS knowledge source server: an object model for delivering UMLS data. AMIA Annu Symp Proc 2003:51–5. [PubMed: 14728132]
44. http://trec.nist.gov/pubs/trec7/t7_proceedings.html
45. Zar, JH. Biostatistical analysis. Prentice Hall Inc; EngleWood Cliffs, NJ: 1974.
46. Haynes RB, Wilczynski N, McKibbon KA, et al. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. J Am Med Inform Assoc 1994;1(6):447–58. [PubMed: 7850570]
47. Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis CF. Text categorization models for high-quality article retrieval in internal medicine. J Am Med Inform Assoc 2005;12(2):207–16. [PubMed: 15561789]
48. Kilicoglu H, Demner-Fushman D, Rindflesch T, Wilczynski NL, Haynes RB. Toward automatic recognition of high quality clinical evidence. Submitted to AMIA Fall Symposium. 2008

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

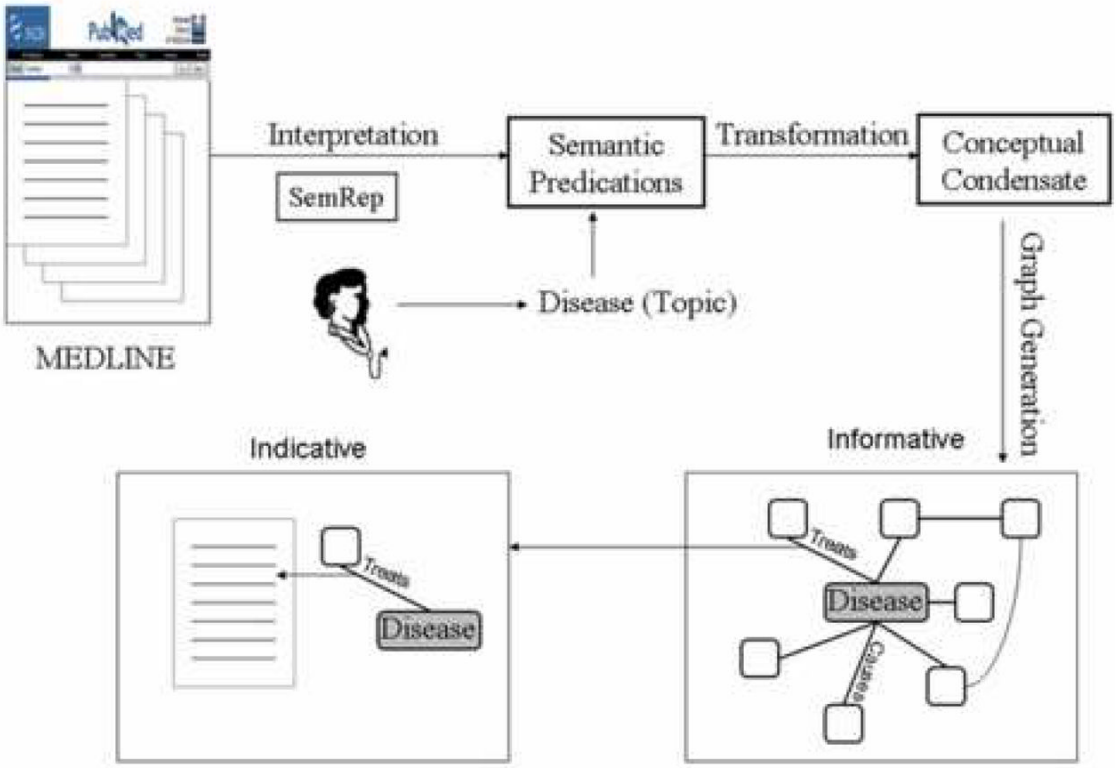

Figure 1.
A schematic view of our semantic abstraction summarizer.

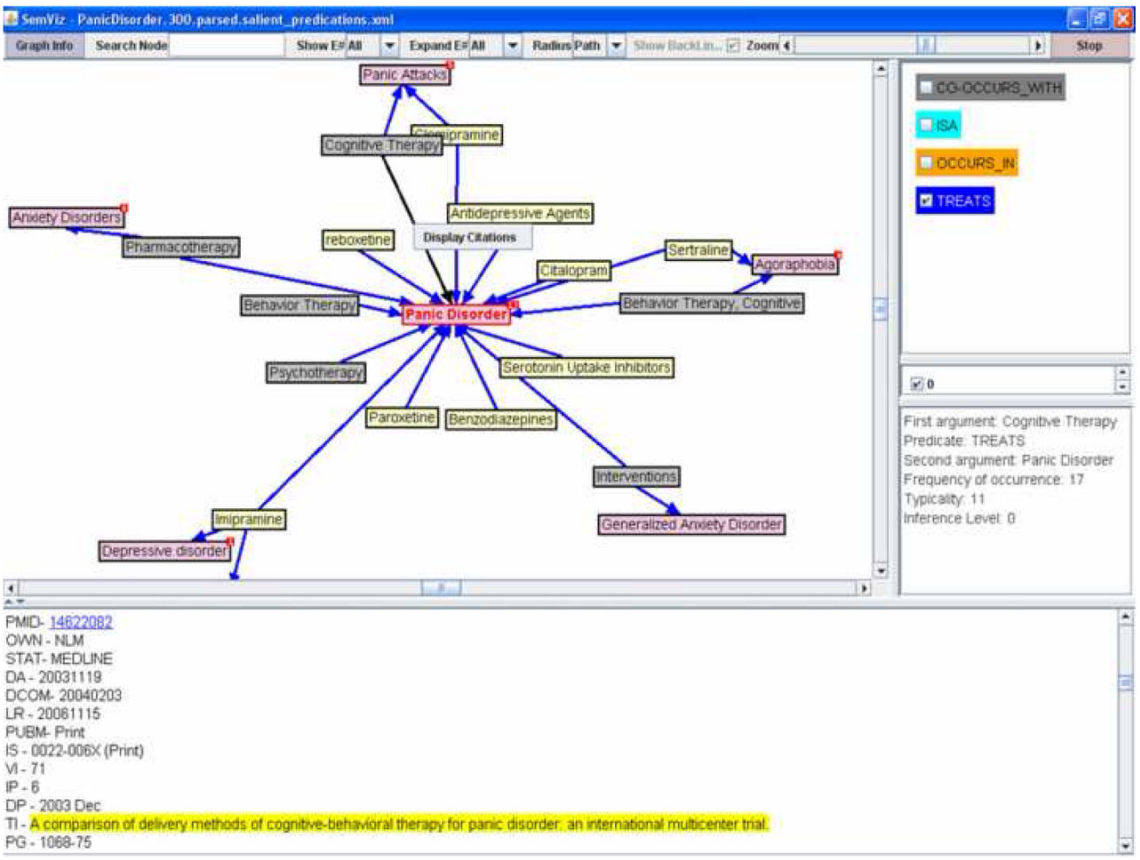

Figure 2.
Partial display of the results of summarizing 300 MEDLINE citations returned by a PubMed search on panic disorder (only interventions or TREATS predications are shown).

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

(1) $term[mh noexp] AND $rest_of_the_words[tw] AND drug therapy[sh]

(2) AND hasabstract[text] AND Clinical Trial[pt] AND English[Lang]

(3) AND humans[mh] AND (1900[PDAT] $year/$month[PDAT])

Figure 3.
PubMed search template where $term, $rest_of_the_words, $year, and $month denote respectively a MeSH term, words not mapped to MeSH, a year and a month in which a CE summary was created.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

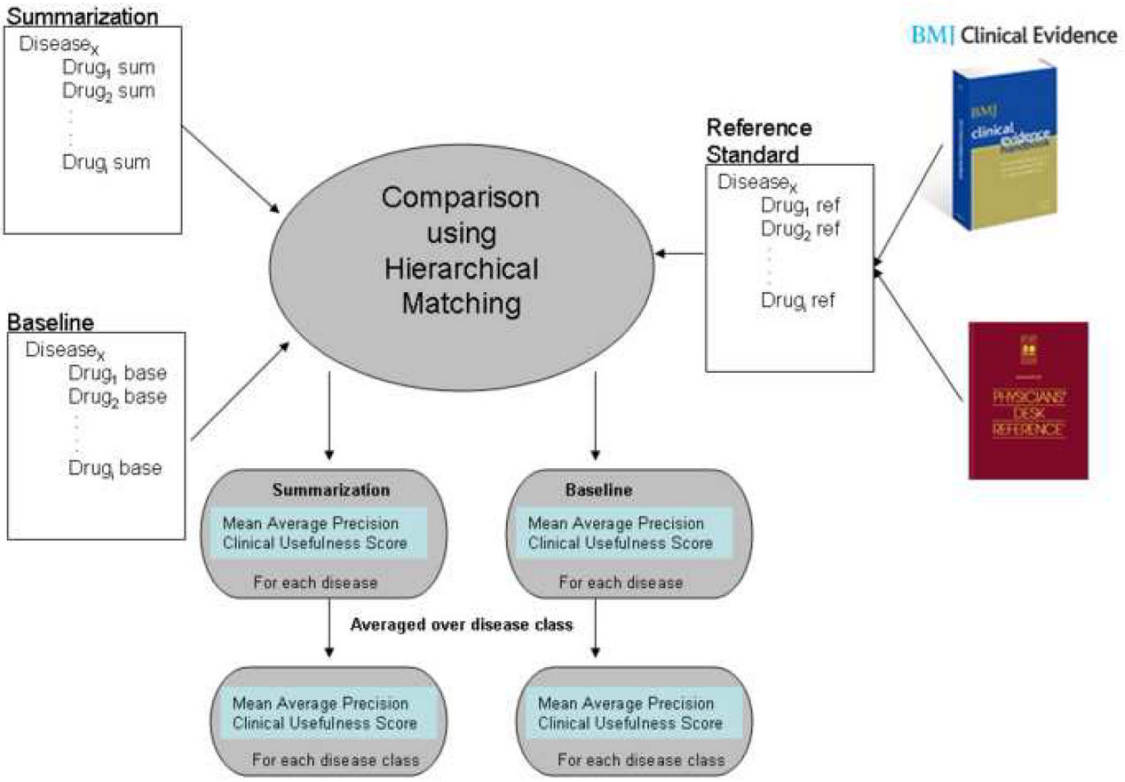

Figure 4.
An overall view of the evaluation methodology.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 1

Topics included in this study. Diseases and classes are defined in Clinical Evidenceconcise.

Disease Class	Topic
Blood and lymph disorders	early stage aggressive non-Hodgkin's lymphoma
Cardiovascular disorders	acute myocardial infarction, acute ischemic stroke, atrial fibrillation, heart failure, unstable angina
Child health	acute otitis media
Digestive system disorders	gastroesophageal reflux disease, irritable bowel syndrome
Ear, nose, and throat disorders	acute sinusitis, otitis externa, seasonal allergic rhinitis
Endocrine disorders	hypothyroidism, obesity
Eye disorders	bacterial conjunctivitis, glaucoma
HIV and AIDS	pneumocystis carinii pneumonia
Infectious diseases	chickenpox, congenital toxoplasmosis, acute diarrhea, malaria, bacterial meningitis
Men's health	chronic prostatitis, erectile dysfunction
Mental health	dementia, depressive disorders, generalized anxiety disorder, mania, panic disorder, post-traumatic stress disorder, schizophrenia
Musculoskeletal disorders	acute low back pain, leg cramps, osteoarthritis, osteoporosis, rheumatoid arthritis
Neurological disorders	epilepsy, migraine, multiple sclerosis, Parkinson's Disease, trigeminal neuralgia
Oral health	oropharyngeal candidiasis
Respiratory disorders (acute)	acute asthma, acute bronchitis, community acquired pneumonia
Respiratory disorders (chronic)	chronic asthma, chronic obstructive pulmonary disease
Sexual health	chlamydia infection, genital warts, gonorrhea
Skin disorders	atopic eczema, chronic psoriasis, scabies, tinea pedis

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 2

Reference standard entry for panic disorder.

CE/PDR Intervention Name	UMLS Concept	Treatment Category
selective serotonin reuptake inhibitors	C0162758: Serotonin Uptake Inhibitors	beneficial
tricyclic antidepressants	C0003290: Antidepressive Agents, Tricyclic	beneficial
imipramine	C0020934: Imipramine	beneficial
benzodiazepines	C0005064: Benzodiazepines	trade-off
buspirone	C0006462: buspirone	unknown
monoamine oxidase inhibitors	C0026457: Monoamine Oxidase Inhibitors	unknown
paroxetine hydrochloride	C0070122: Paroxetine C0771019: Paroxetine Hydrochloride	pdr
sertraline hydrochloride	C0074393: Sertraline C0600526: Sertraline Hydrochloride	pdr
clonazepam	C0009011: Clonazepam	pdr
alprazolam	C0002333: Alprazolam	pdr
fluoxetine hydrochloride	C0733380: Fluoxetine Hydrochloride C0016365: Fluoxetine	pdr

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 3

Appendix

Interventions as Metathesaurus concepts found by the summarization system for each topic included in the study. In the last column, the first number inside the parenthesis is frequency of occurrence and the second number is typicality.

Disease class	Topic	Interventions
Blood and lymph disorders	early stage aggressive non-Hodgkin's lymphoma	NONE
Cardiovascular disorders	acute ischaemic stroke	Alteplase(8/6); Aspirin(5/3); Ancrod(4/2)
	acute myocardial infarction	Thrombolytic Therapy(52/37); Alteplase(28/20); Streptokinase(22/19); Angiotensin-Converting Enzyme Inhibitors(17/13); Reperfusion Therapy(16/13); Fibrinolytic Agents(15/15); Aspirin(11/9); Angioplasty, Transluminal, Percutaneous Coronary(9/6); Heparin(8/7); Captopril(8/6); Angioplasty(7/6); reteplase(7/6); Administration (procedure)(6/6); Metoprolol(5/4); Thrombin(4/4); Unfractionated heparin (substance)(4/2); Ramipril(4/3); Lisinopril(4/2); Heparin, Low-Molecular-Weight(3/2)
	atrial fibrillation	Digoxin(13/7); Diltiazem(8/4); Anticoagulant therapy (procedure)(5/2); Verapamil(2/2); esmolol(2/2); Amiodarone(2/1)
	heart failure	carvedilol(19/12); Adrenergic beta-Antagonists(16/12); Angiotensin-Converting Enzyme Inhibitors(11/10); Metoprolol(10/8); Digoxin(8/5); valsartan(6/4)
	unstable angina	Heparin(27/21); Aspirin(13/12); Unfractionated heparin (substance)(12/9); Propranolol(11/5); Metoprolol(11/5); Alteplase(9/5); Enoxaparin(7/5); Nifedipine(6/4); Heparin, Low-Molecular-Weight(5/5); tirofiban(5/5); Calcium Channel Blockers(3/2); inogatran(3/2); Thrombin(2/2); GENERAL OPERATIVE PROCEDURES(2/2); Hirudin(2/2); Coronary Artery Bypass(2/2)
Child health	acute otitis media	Amoxicillin(51/34); Cefaclor(39/24); Clavulanate(25/13); Antibiotics(22/19); Amoxicillin-Potassium Clavulanate Combination(16/12); Ceftriaxone(16/7)
Digestive system disorders	gastroesophageal reflux disease	Omeprazole(42/29); Cisapride(33/23); Ranitidine(29/21); Metoclopramide(22/15); Cimetidine(14/10); lansoprazole(10/8); Proton pump inhibitor (substance)(8/8); pantoprazole(6/5); Baclofen(5/3); rabeprazole(5/3); Antacids(4/3); Famotidine(3/2); Domperidone(3/1)
	irritable bowel syndrome	Excluded from study. Search did not find citations.
Ear, nose, and throat disorders	acute sinusitis	Amoxicillin(8/5); Doxycycline(7/4); cefuroxime axetil(6/4)
	otitis externa	Otosporin(3/2); Tobrex(2/1)
	seasonal allergic rhinitis	Loratadine(20/13); Cetirizine(20/13); Budesonide(17/13); Antihistamines(14/14); Terfenadine(13/9); Adrenal Cortex Hormones(12/8); Beclomethasone Dipropionate(11/8); Fluticasone propionate(11/10); ebastine(10/6); Cromolyn Sodium(8/4); Triamcinolone Acetonide(8/6); acrivastine(7/3); Fexofenadine hydrochloride(7/3); fexofenadine(7/6); levocabastine(6/3); Azelastine(6/5); Pseudoephedrine(5/4); Astemizole(4/4); The science and art of healing(3/3); Beclomethasone(3/3)
Endocrine disorders	hypothyroidism	Thyroxine(11/9); Replantation(2/2); Triiodothyronine(2/2); Thyroxine therapy(2/2)
	obesity	sibutramine(25/15); orlistat(24/22); Fluoxetine(20/14); Dexfenfluramine(14/11); Metformin(12/11); Leptin(7/3); Insulin(7/7); troglitazone(5/4); Gemfibrozil(4/2); Dietary Supplementation(3/3); Calorie restricted diet (finding)(3/3)
Eye disorders	bacterial conjunctivitis	Eyedrops(2/2); Ciprofloxacin(2/2); Ofloxacin(2/1)
	glaucoma	Timolol(13/9); Adrenergic beta-Antagonists(13/8); brimonidine(7/5); Epinephrine(5/5); Pilocarpine(4/4); Trabeculectomy(4/4); Eyedrops(4/4); Mitomycin(4/4); 4-benzamido-4′-isothiocyanostilbene-2,2′-disulfonate(3/2); Betaxolol(3/3)
HIV and AIDS	pneumocystis carinii pneumonia	Trimethoprim-Sulfamethoxazole Combination(24/20); Pentamidine(11/9); Prophylactic treatment(10/8); Adrenal Cortex Hormones(7/6); Dapsone(4/2); hydroxynaphthoquinone 566C80(3/2)
Infectious diseases	acute diarrhea	Octreotide(17/10); Loperamide(7/4); Octreotide Acetate(6/4); nitazoxanide(4/3); Fluorouracil(4/3); Chemotherapy-Oncologic Procedure(3/2); alosetron(3/2); Radiation therapy(2/2); Antibiotics(2/2);

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Disease class	Topic	Interventions
		Roxithromycin(2/1); Antidiarrheals(2/1); Administration, Oral(2/1); Thalidomide(2/2); Smectite(2/1); Albendazole(2/1); Metronidazole(2/2)
	bacterial meningitis	Ceftriaxone(19/12); Chloramphenicol(8/7); Cefotaxime(7/6); Ampicillin(5/5)
	chickenpox	Acyclovir 8/13)
	congenital toxoplasmosis	NONE
	malaria	Chloroquine(5/5); Mefloquine(4/3); halofantrine(3/1); sulfadoxine-pyrimethamine(3/2)
Men's health	chronic prostatitis	Ciprofloxacin(6/5)
	erectile dysfunction	sildenafil(96/55); Alprostadil(57/36); Viagra(29/24); Sildenafil citrate(28/21); Apomorphine(9/5); Phentolamine(7/4); Injections(6/6); Intracavernous injection (procedure)(6/6); Papaverine(5/4); Glonoine, nitroglycerine, Homeopathic preparation(4/2)
Mental health	dementia	Risperidone(5/4); olanzapine(4/3); Antipsychotic Agents(4/2); donepezil(4/3); Ginkgo biloba extract(4/2); Melatonin(2/1); The science and art of healing(2/2); Levodopa(2/2)
	depressive disorder	Fluoxetine(23/18); Antidepressive Agents(17/17); Paroxetine(10/9); Serotonin Uptake Inhibitors(10/9); Imipramine(9/5); Release procedure(7/6); nefazodone(7/4); reboxetine(6/2); Sertraline(6/4); Fluvoxamine(6/4); Nortriptyline(5/4); venlafaxine(5/4); Amitriptyline(4/3); Psychotherapy(4/3); Tryptophan(3/3); Mirtazapine(3/2); Augmentation procedure (procedure)(3/3); Desipramine(2/2); Testosterone(2/1); olanzapine(2/2); Citalopram(2/2); Interventions(2/2); Pharmacotherapy(2/2); Bupropion(2/2); Electroconvulsive Therapy(2/2); Administration (procedure)(2/1)
	generalized anxiety disorder	NONE
	mania	Lithium(16/13); Verapamil(8/6); Valproate(6/5); Carbamazepine(6/6); Haldol decanoate, homeopathic preparation(6/6); Antipsychotic Agents(5/5); olanzapine(4/4); The science and art of healing(4/4); Clozapine(4/2); Antimanic Agents(4/4); Electroconvulsive Therapy(3/2); Risperidone(3/3)
	panic disorder	Imipramine(31/23); Clonazepam(18/12); Fluoxetine(17/11); Paroxetine(15/10); Benzodiazepines(15/8); Sertraline(14/8); Alprazolam(14/12); Serotonin Uptake Inhibitors(12/12); Clomipramine(12/11); Cognitive Therapy(10/7); Fluvoxamine(9/6); Antidepressive Agents(8/6); Behavior Therapy, Cognitive(8/4); venlafaxine(7/3); Interventions(4/2)
	post-traumatic stress disorder	Sertraline(15/7); Fluoxetine(12/8); nefazodone(11/5); Mirtazapine(6/3); Paroxetine(5/3)
	schizophrenia	Antipsychotic Agents(41/32); Clozapine(37/28); Haldol decanoate, homeopathic preparation(35/28); olanzapine(30/24); Risperidone(29/23); Atypical antipsychotic (substance)(8/8); The science and art of healing(3/3)
Musculoskeletal disorders	acute low back pain	tenoxicam(4/1)
	leg cramps	NONE
	osteoarthritis	Naproxen(45/35); Anti-Inflammatory Agents, Non-Steroidal(34/27); Piroxicam(33/27); Diclofenac(31/24); nabumetone(24/18); Indomethacin(19/15); Ibuprofen(17/13); Etodolac(12/10); Anti-Inflammatory Agents(8/7); Aspirin(7/4); meloxicam(7/5); Hyaluronic Acid(6/5); tenoxicam(5/5); Acetaminophen(5/4); Cyclooxygenase-2 Inhibitors(4/3); Misoprostol(4/4); Sulindac(4/3)
	osteoporosis	Alendronate(12/9); Etidronate(9/7); Natrium fluoratum, sodium fluoride, Homeopathic preparation(7/6); amidronate(7/4); Fluorides(6/4); Raloxifene(5/2); Vitamin D(4/2); Calcitriol(3/3); Parathyroid Hormones(2/2); Calcitonin(2/2); salmon calcitonin(2/2); Methandrostenolone(2/2); Calcium supplementation(2/2); Estrogens(2/2)
	rheumatoid arthritis	Methotrexate(108/72); Cyclosporine(27/18); Sulfasalazine(21/16); leflunomide(20/14); Disease-Modifying Second-Line Drugs(17/11); The science and art of healing(16/15); Combination electrotherapy(15/12); interleukin-1 receptor antagonist(11/5); Minocycline(11/6); Diclofenac(9/8); meloxicam(8/5); Auranofin(8/7); Adrenal Cortex Hormones(7/7); Cyclophosphamide(7/5); Anti-Inflammatory Agents, Non-Steroidal(6/6); Administration (procedure)(6/5); Glucocorticoids(5/4); Antirheumatic Agents(5/5); Anti-Inflammatory Agents(5/5); Hydroxychloroquine(5/4); aceclofenac(5/3); Erythropoietin, Recombinant(4/2)
Neurological disorders	epilepsy	Vigabatrin(29/22); Carbamazepine(24/20); Antiepileptic Agents(23/19); lamotrigine(21/17); Phenytoin(14/10); Valproate(13/9); Valproic Acid(10/8); clobazam(6/3); Phenobarbital(5/5); topiramate(5/5); felbamate(5/3); gabapentin(5/4); tiagabine(4/3); Flunarizine(3/3); Anticonvulsants(3/3)
	migraine	Sumatriptan(129/87); Metoclopramide(17/13); rizatriptan(14/9); zolmitriptan(13/7); almotriptan(12/8); Dihydroergotamine(10/8); frovatriptan(9/4); 311C90(9/5); Acetaminophen(9/7); Prochlorperazine(8/4);

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Disease class	Topic	Interventions
	multiple sclerosis	naratriptan(7/5); Flunarizine(7/6); eletriptan(7/6); Valproate(6/2); Caffeine(6/4); Analgesics(5/5); Ibuprofen(5/3); Propranolol(4/4); Sodium Valproate(4/4); Administration (procedure)(4/4); avitriptan(4/2); Anti-Inflammatory Agents, Non-Steroidal(3/3); Meperidine(3/2) Methylprednisolone(23/18); Azathioprine(13/8); interferon beta 1a(11/7); Interferon-beta(9/6); interferon beta-1b(7/5); Corticotropin(6/4); Therapeutic immunosuppression(6/6); gabapentin(6/5); Cyclosporine(5/5); Baclofen(5/5); Amantadine(5/3); Interferon-alpha(5/4); Cyclophosphamide(4/3); tizanidine(4/2); Copaxone(4/4); Isoniazid(4/3); The science and art of healing(4/4); Avonex(4/4)
	Parkinson Disease	Levodopa(79/62); Bromocriptine(20/16); Dopamine Agonists(19/15); Apomorphine(17/13); pramipexol(16/11); ropinirole(15/12); Selegiline(15/10); tolcapone(9/8); Deprenil(9/5); Pergolide(8/6); Stimulation - action (qualifier value)(8/7); entacapone(8/6); cabergoline(8/6); budipine(8/3); Carbidopa(6/5); Amantadine(5/3); olanzapine(4/4); Anticholinergic Agents(3/2)
	trigeminal neuralgia	Baclofen(7/4); Carbamazepine(3/3); Capsaicin(2/2); Anticonvulsants(2/1)
Oral health	oropharyngeal candidiasis	Clotrimazole(10/8); Miconazole(8/4); Fluconazole(8/5); Nystatin(7/5)
Respiratory disorders (acute)	acute asthma	Albuterol(44/31); Adrenal Cortex Hormones(12/12); Epinephrine(11/9)
	acute bronchitis	Erythromycin(10/7); Amoxicillin(7/4); Antibiotics(6/4); Cefaclor(6/4); cefuroxime axetil(4/3); Azithromycin(4/2); dirithromycin(3/2); Clavulanate(3/2); Amoxicillin-Potassium Clavulanate Combination(2/1)
	community acquired pneumonia	Azithromycin(10/4); Clarithromycin(7/4); ANTIMICROBIALS(4/4); Erythromycin(4/3); Ceftriaxone(4/3); Antibiotics(3/2); Amoxicillin(3/2); Ofloxacin(3/2); Cefonicid(2/1)
Respiratory disorders (chronic)	chronic asthma	Albuterol(34/31); Adrenal Cortex Hormones(32/29); Theophylline(28/22); Beclomethasone Dipropionate(22/16); Bronchodilator Agents(19/17); Budesonide(17/12); Leukotriene Antagonists(15/10); montelukast(12/12); salmeterol(9/7); Beclomethasone(9/8); Cyclosporine(8/5); Fenoterol(8/7); Fluticasone propionate(7/7); Nedocromil, Disodium Salt(7/5); Ipratropium Bromide(6/6); flunisolide(5/3); Terbutaline(5/5); Anti-Inflammatory Agents(3/3); The science and art of healing(3/3)
	chronic obstructive pulmonary disease	salmeterol(10/8); Adrenal Cortex Hormones(7/6); Bronchodilator Agents(7/5); Ipratropium Bromide(6/4); Fluticasone propionate(6/5); Theophylline(6/4); formoterol(6/3); tiotropium(5/3); Almitrine Bismesylate(3/1)
Sexual health	chlamydia infection, genital warts	Doxycycline(9/7); Ofloxacin(8/6); Azithromycin(6/5); Erythromycin(5/5) Podophyllotoxin(20/12); imiquimod(8/4); Isotretinoin(7/3); Emollient Cream(6/5); Fluorouracil(6/3); Podophyllin(4/3)
	gonorrhea	Ceftriaxone(41/30); Probenecid(25/20); Ciprofloxacin(19/12); Ofloxacin(17/11); Penicillin(13/12); Spectinomycin(13/13); cefuroxime axetil(11/6); Cefotaxime(11/7); Amoxicillin(9/7); Cefixime(9/7); Trimethoprim-Sulfamethoxazole Combination(7/5); Aztreonam(7/4); Azithromycin(7/4); Enoxacin(7/5); Penicillin G, Procaine(6/6); Quinolones(5/5); Augmentin(5/3); Thiamphenicol(4/3); Ampicillin(4/4); Antibiotics(3/3); Penicillin G(3/3)
Skin disorders	atopic eczema	Cyclosporine(41/21); Tacrolimus(23/13); Emollient Cream(22/19); The science and art of healing(7/5); gamma-Linolenic Acid(4/3); Antihistamines(3/2); Betamethasone valerate(3/3);
	chronic psoriasis	Cyclosporine(17/13); calcipotriene(6/5); Anthralin(5/5); Tacalcitol(4/3); PUVA(4/4); alefacept(3/2); Retinoids(3/3);
	scabies	Ivermectin(19/9);
	tinea pedis	Emollient Cream(24/15); terbinafine(15/10); Griseofulvin(7/4); Itraconazole(7/4);

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 4

Interventions for the topic dementia. Those considered "beneficial" or "likely to be beneficial" in the reference standard are underlined.

Summarization system	Baseline
Donepezil	Pharmaceutical Preparations
Ginkgo biloba extract	Antipsychotic agents
Antipsychotic Agents	Risperidone
Risperidone	Haloperidol
Olanzapine	Olanzapine

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 5

Mean average precision scores for disease classes for baseline (BASE) and the summarization system (SUM). In the first column, N is the number of topics in a disease class. The last column is the gain in MAP. +Statistical significance ($p < 0.01$).

Disease Class	Topic (N)	BASE	SUM	MAP Gain
Oral health	1	0.10	1.00	0.90
Digestive system disorders	1	0.06	0.71	0.65
Respiratory disorders (chronic)	2	0.15	0.78	0.63
Musculoskeletal disorders	5	0.36	0.78	0.43
Respiratory disorders (acute)	3	0.08	0.47	0.38
Endocrine disorders	2	0.50	0.83	0.33
Mental health	7	0.49	0.69	0.20
Cardiovascular disorders	5	0.50	0.60	0.11
Ear, nose, and throat disorders	3	0.21	0.31	0.10
Sexual health	3	0.42	0.52	0.09
Men's health	2	0.38	0.46	0.07
Skin disorders	4	0.22	0.23	0.01
Neurological disorders	5	0.41	0.42	0.01
Infectious diseases	5	0.28	0.28	0.00
Blood and lymph disorders	1	0.00	0.00	0.00
Child health	1	0.03	0.00	−0.03
Eye disorders	2	0.08	0.00	−0.08
HIV and AIDS	1	0.84	0.51	−0.33
All	**53**	**0.33**	**0.50**	**0.17** +

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 6

Overall score of clinical usefulness for the baseline(BASE) and summarization system (SUM). The last column is the gain in the overall clinical usefulness score. + Statistical significance ($p < 0.05$).

Disease Class	Topic (N)	BASE	SUM	Overall-Gain
Oral health	1	−0.20	2.50	2.70
Respiratory disorders (acute)	3	−0.33	1.00	1.33
Respiratory disorders (chronic)	2	0.00	0.86	0.86
Digestive system disorders	1	−0.60	0.23	0.83
Mental health	7	0.49	1.11	0.62
Musculoskeletal disorders	5	0.36	0.89	0.53
Infectious diseases	5	−0.36	0.04	0.40
Skin disorders	2	−0.10	0.16	0.26
Ear, nose, and throat disorders	4	0.73	0.92	0.19
Eye disorders	3	−0.40	−0.23	0.17
Men's health	2	0.00	0.09	0.09
Child health	5	1.00	1.00	0.00
Neurological disorders	1	0.36	0.36	0.00
Blood and lymph disorders	1	−1.00	−1.00	0.00
Sexual health	3	0.87	0.69	−0.18
Cardiovascular disorders	5	0.84	0.64	−0.20
Endocrine disorders	2	0.00	−0.33	−0.33
HIV and AIDS	1	2.20	1.67	−0.53
All	**53**	**0.25**	**0.64**	**0.39**$^{+}$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 7

Manual comparison between summarization and the reference standard

Topic	Recall	Precision
Acute myocardial infarction	95%	94%
Gastroesophageal reflux disease	100%	54%
Community-acquired pneumonia	100%	78%
Panic disorder	100%	57%
Overall	**98%**	**73%**