Published in final edited form as:
Stud Health Technol Inform. 2013 ; 192: 846–850.

Automatic identification of comparative effectiveness research from Medline citations to support clinicians' treatment information needs

Mingyuan Zhang[a], **Guilherme Del Fiol**[a], **Randall W. Grout**[b], **Siddhartha Jonnalagadda**[c], **Richard Medlin Jr**[d], **Rashmi Mishra**[a], **Charlene Weir**[a], **Hongfang Liu**[c], **Javed Mostafa**[d], and **Marcelo Fiszman**[e]

[a]Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA

[b]Carver College of Medicine, University of Iowa, Iowa City, IA, USA

[c]Natural Language Processing Group, Mayo Clinic, Rochester, MN, USA

[d]School of Information and Library Science, University of North Carolina, Chapel Hill, NC, USA

[e]Lister Hill Center, National Library of Medicine, Bethesda, MD, USA

Abstract

Online knowledge resources such as Medline can address most clinicians' patient care information needs. Yet, significant barriers, notably lack of time, limit the use of these sources at the point of care. The most common information needs raised by clinicians are treatment-related. Comparative effectiveness studies allow clinicians to consider multiple treatment alternatives for a particular problem. Still, solutions are needed to enable efficient and effective consumption of comparative effectiveness research at the point of care.

Objective—Design and assess an algorithm for automatically identifying comparative effectiveness studies and extracting the interventions investigated in these studies.

Methods—The algorithm combines semantic natural language processing, Medline citation metadata, and machine learning techniques. We assessed the algorithm in a case study of treatment alternatives for depression.

Results—Both precision and recall for identifying comparative studies was 0.83. A total of 86% of the interventions extracted perfectly or partially matched the gold standard.

Conclusion—Overall, the algorithm achieved reasonable performance. The method provides building blocks for the automatic summarization of comparative effectiveness research to inform point of care decision-making.

Keywords

comparative effectiveness research; information needs

Introduction

In the course of clinical practice, health care professionals often raise needs for information to support patient care decision-making. These questions are frequently left unanswered due

Address for correspondence: Mingyuan Zhang, 26 South 2000 East Room 5775 HSEB, Salt Lake City, UT 84112-5775, mingyuan.zhang@utah.edu.

to lack of readily available knowledge resources and limited time to access these resources in the health care workflow. Previous studies have demonstrated that providing physicians linkage to knowledge resources at the point of care can effectively address these needs and encourage evidence-based practice.[1]

Comparing available treatments for a particular problem is one of the most frequent types of patient care information needs.[2] Comparative effectiveness studies are designed to answer these kinds of questions.[3] Online resources such as Medline provide access to answers to most patient care clinical questions.[4] Although comparative effectiveness studies are indexed in Medline, to apply this kind of research to a particular patient, clinicians may need to scan through several studies. Systems that automatically summarize the state-of-the-art in a given topic are promising solutions for efficient and effective consumption of comparative effectiveness research at the point of care.

The goal of this study was to design and assess an algorithm for automatically identifying comparative effectiveness studies on the treatment of a given condition and extracting the interventions investigated in these studies. The study hypotheses were: 1) The algorithm can accurately identify comparative effectiveness studies in Medline; 2) the algorithm can accurately extract the treatment interventions that are compared in these studies.

Background

A study by Ely et al. indicated that roughly 40% of physicians' questions are related to treatment.[2] For this kind of information need, clinicians could benefit from studies that directly compare the effectiveness, safety, and tolerability of multiple health care interventions to decide which one is optimal for a particular patient. Comparative effectiveness studies provide this type of information. Unlike the rigorously controlled environments in placebo randomized clinical trials, comparative effectiveness studies directly compare multiple treatment alternatives in typical practice.[3] Synthesizing the results of comparative effectiveness studies is important for enabling their consumption in the patient care decision-making process.

Previous work

In a previous study, we assessed the feasibility of generating knowledge summaries composed of relevant sentences extracted from Medline citations.[5] The knowledge summary system consists of a pipeline that integrates multiple natural language processing (NLP) tools and information retrieval resources, including the UMLS Metathesaurus for concept extraction and SemRep for semantic predication extraction.[6] In two case studies, the system achieved a high precision in extracting sentences related to depression and Alzheimer's disease (91.3%), but only 10% of the sentences retrieved compared two or more treatment alternatives.[5] In the present study, we aimed at enhancing our algorithm with focus on comparative studies. More specifically, we focused on two necessary steps for this kind of summarization: identifying comparative effectiveness citations and extracting the treatment interventions being compared. This kind of process enables summarization techniques such as semantic grouping.[7] For example, sentences on the "treatment of nocturnal enuresis in children" could be grouped according to different treatment approaches: behavior therapy, alarm intervention, and tricyclic antidepressants.

Our method combines semantic predications from the Semantic MEDLINE database (SemMedDB), Medline citation metadata, and machine learning techniques. The tools and resources used in our algorithm are described below.

Semantic MEDLINE database (SemMedDB)

Semantic Medline is a semantic knowledge summarization tool that summarizes Medline citations on a particular topic. SemMedDB stores semantic predications extracted from Medline citations. The semantic predications consist of relations between UMLS concept arguments, e.g., “X treats Y” where X and Y are UMLS concepts.[6] The database is populated by a linguistic parser called SemRep, which uses underspecified syntactic analysis and structured domain knowledge from the UMLS. There are several semantic predication types in the database that interpret comparative constructs in the biomedical literature, including “same_as”, “lower_than”, “higher_than” and “compared_with”.[8] Table 1 provides examples of these constructions. SemMedDB predications have a “novelty” attribute, which indicates whether a predication is interesting or too generic. Novelty is determined by the contextual distance between a concept argument to the root of the concept hierarchy.[9] Concepts that are too close to the root are considered not to be novel.

Medline metadata elements

Medline records contain an extensive set of metadata.[10] The following metadata fields are relevant to this study: 1) “Publication Type,” which may take several values, including “Comparative Study”; 2) “Registry Number (RN),” which contains a 5 to 9-digit code assigned by the Chemical Abstract Service (CAS) to identify chemical substances that are subject of investigation; and 3) “Name of Substance (NM),” which contains a human-readable term of an RN code. NM terms and RN codes are included in the Medical Subject Headings (MeSH) and the MeSH Supplementary Concept Records. A value of “0” is assigned to the RN field in the case of drug classes and substances that are not available in CAS. Figure 1 shows a citation that includes the RN and NM fields.

Method

The study method consisted of: 1) developing an algorithm that integrates SemMedDB and Medline metadata to identify comparative studies and extract interventions; 2) developing a gold standard using the treatment of depression as a case study; and 3) evaluating the algorithm performance.

Algorithm Description

The algorithm consisted of 5 steps: 1) Retrieval of comparative citations; 2) extraction of comparative and treatment predications from SemMedDB; 3) extraction of Medline citation fields; 4) a comparative study classifier; and 5) identification of study interventions. Algorithm design was guided by manual analysis of 55 randomly selected citations from the gold standard. Figure 2 depicts the algorithm steps.

Step 1- Retrieval of comparative citations from Medline—Candidate citations were retrieved using the knowledge summary system developed by our previous study. The system uses a set of heuristics that include optimal search strategies for retrieving clinically useful articles. Details of this algorithm are provided elsewhere.[5]

Step 2 – Extraction of comparative and treatment predications from SemMedDB—We queried SemMedDB to retrieve all the predications with a treatment or comparative predication (i.e., “treats,” “same_as,” “lower_than,” “higher_than,” “compared_with”) and an object related to depression. Candidate interventions were extracted from the subjects of the treatment predications and both subjects and objects of the comparison predications. Non-novel predications were not included. In addition, a set of uninteresting or overly generic treatment arguments, such as “Pharmacotherapy” and “Intervention regimes,” were identified and removed. Predications in which the object or subject was “Placebo” were also removed since we focused only on non-placebo comparisons.

Next, redundant predications due to drug/drug class relationships were removed. For example, “Trimipramine” and “Antidepressive Agents, Tricyclic” were identified as treatments of depression in the same citation. Since “Trimipramine” is a child of “Antidepressive Agents, Tricyclic,” the latter was removed. To identify drug/drug class relationships, we implemented a program that uses the UMLS Metathesaurus MRHIER and MRREL tables. The MRHIER table contains hierarchy information from a given concept to the root concept, thus providing a complete list of antecessors for a given concept. The MRREL table provides information on relations between concepts, such as “PAR” (parent relationship) and “RB” (child-parent relationship). If concept A is on the concept B’s path to the root concept or concept A has a “PAR” or “RB” relationship with concept B, concept A is determined to be the drug class of concept B and therefore is removed from the list of study interventions.

Step 3 – Extraction of Medline citation fields—In this step, we utilized the NLM Entrez Programming Utilities to retrieve Medline citations in XML format using the PMIDs of the candidate citations. The RN and NM fields were parsed out by a Java DOM XML parser. Using the MeSH Headings and Supplementary Concept Records, we extracted the following MeSH attributes for each RN field: MeSH code, MeSH concept name, UMLS CUI, preferred term flag, and UMLS concept name. The drugs compared in the studies were identified as those represented in the RN instances. In the case of redundant RN instances (i.e., both a drug and its drug class had an RN entry), we removed the drug class by using the algorithm described above in Step 2. In addition, we attempted to exclude RN instances that are not used to treat the condition of interest (i.e., depression), such as “Kynurenine.” For this process, we checked if a given RN instance was listed as a treatment of depression in NDF-RT (“may treat” relation) or SemMedDB (“treats” predications). RN instances that were not included as a treatment of depression on either source were removed from the output.

Step 4 – Comparative study classifier—In this step, we developed a machine learning classifier to identify comparative studies. The dataset consisted of predictors extracted in the previous steps for each of the citations in the depression gold standard. The following predictors were included: 1) whether the Medline “Publication type” field indicates a “Comparative Study”; 2) number of interventions extracted from the RN field; 3) number of interventions extracted from comparative predications; 4) number of interventions extracted from “treats” predications; and 5) number of different interventions regardless of source. Using the training set described in the “Gold Standard” section, we produced classification models using 5 techniques available in the Weka machine learning framework [11]: Naïve Bayes, Bayesian network, rules (PART algorithm), decision tree (J48 algorithm) and support vector machine (SVM). The performances of these candidate classifiers were assessed using the independent testing set.

Step 5 – Identification of study interventions—To identify the interventions compared in each study, we simply merged the list of interventions that were extracted and cleaned from SemMedDB and Medline RN fields. In the merging process, duplicates were removed.

Gold Standard

The gold standard consisted of 351 Medline citations on the treatment of depression (Figure 3). These citations were retrieved using a pre-defined search strategy that relies on

PubMed's Clinical Queries filter. This filter is tuned to retrieve high quality clinical articles on a given topic. Details of this algorithm are available elsewhere.[5] We annotated the retrieved citations according to the following attributes: 1) whether the study was relevant to the topic at hand (i.e., treatment of depression); 2) whether the study compared two or more treatment interventions; 3) the intervention comparison pairs; and 4) the direction of the comparison (e.g., treatment A *higher than* treatment B).

The gold standard was developed iteratively. In the first step, a random sample of 20 citations was analyzed to develop annotation guidelines as follows: 1) if the citation is not relevant to the topic, do not rate the study as comparative or not; 2) if the citation is rated as a comparative study, raters should also annotate the interventions being compared, and the direction of the comparison. Next another random set of 20 citations was annotated independently by two clinicians (Medlin and Mishra) for calibration. Since the inter-rater agreement (Kappa) in the second set was strong for all annotation attributes (0.83 for relevancy, 0.62 for comparative study, and 1.0 for interventions), the remaining 311 citations were annotated by one of the two clinicians (Mishra).

Out of the 351 citations retrieved, 256 were relevant and 110 were rated as comparative studies (Figure 3). We randomly selected 37 out of the 110 comparative citations to guide the development of the intervention extraction algorithm (Figure 2, Step 5). In addition, we randomly selected two thirds of the 256 citations for training and one third for testing the comparative study classifier (Figure 2, Step 4).

Evaluation—The treatments automatically extracted were compared with the gold standard. The following measures were obtained: 1) Precision, recall and F-measure of the Publication Type field for identifying comparative studies; 2) precision, recall, F-measure, and the area under the ROC curve (AUC) of the comparative study classifier; 3) percentage of comparative citations from which the algorithm completely or partially identified the study interventions; and 4) percentage of comparative citations from which the algorithm correctly identified the comparison direction.

Results

Table 2 summarizes the performance of identifying comparative studies using the Publication Type field only and with each of the classifiers. For intervention extraction, the algorithm output completely agreed with the gold standard in 41 out of 73 (56.2%) comparative citations and at least partially agreed with the gold standard in 63 (86.2%) citations. The method that relied on SemMedDB to identify comparison directions yielded a recall of 6.8% and a precision of 45.5%.

Discussion

In this study, we designed and assessed an algorithm that identifies comparative effectiveness treatment studies in Medline and extracts the interventions being compared in these studies. These steps can be part of a broader pipeline that automatically summarizes comparative effectiveness research to support point of care decision-making. For example, once studies and interventions are identified, the most relevant sentences can be extracted to produce a narrative summary. In addition, a summary can group sentences according to the types of interventions being investigated.

All five machine learning classifiers performed well in identifying comparative studies. Although the study was not designed to determine which classification algorithm works best, it allowed us to conclude that the overall approach is a promising alternative to identifying comparative studies in Medline by using the Publication Type field alone, which had relatively low recall. For example, a citation that compares fluvoxamine with mianserin on depression treatment is indexed in Medline as a randomized clinical trial, but not as a comparative study. Our algorithm correctly identify this citation as a comparative study.

Our algorithm includes a data cleaning process in which predications that have uninteresting predication arguments, such as “Pharmacotherapy,” are removed. Part of this process is achieved using the SemMedDB novelty attribute. Yet, we still had to manually compile a list of concepts that were tagged as novel in SemMedDB, but did not provide useful information for our purpose.

The algorithm also performed well when extracting interventions in comparative studies. A failure analysis revealed that most of the cases incorrectly processed by the algorithm fell into the following three categories: 1) Comparisons that involve a combination therapy in one or more study arms. For example, tryptophan-nicotinamide versus tryptophan-nicotinamide-imipramine. In these cases, the algorithm was often able to identify the individual components, but not which of these components were combined in a treatment regimen. 2) Comparison between different forms of the same drug/procedure or different doses of the same drug (e.g., standard release versus controlled-release, standard Repetitive Transcranial Magnetic Stimulation (rTMS) versus EEG-based rTMS, 30mg versus 60mg). In these cases, the algorithm was able to identify the drug ingredient but not the specific form or dose information. These types of comparisons are a known challenge to SemRep.[9] 3) Non-pharmaceutical interventions (e.g., face-to-face versus online therapy). These cases were more difficult since SemMedDB often contains more general concepts for non-drug interventions (e.g., behavior therapy) and the Medline RN field does not include non-pharmaceutical interventions.

Alternative approaches are needed to address the cases above. For example, MeSH Headings such as “Drug Therapy, Combination” may indicate that a study includes some sort of combination therapy. A second approach is to explore intervention and study outcome information in clinical trial registries such as ClinicalTrials.gov. In addition, Chung developed a method to identify intervention arms in clinical trials using coordinating constructions.[12] Our work is complementary since predications in SemMedDB reflect noun phrase coordination in sentences. For example, the sentence “escitalopram or sertraline for the treatment of depression” would be represented as “escitalopram TREATS depressive disorder”; and “sertraline TREATS depressive disorder.” We plan to take advantage of coordination structures in the future.

The algorithm did not perform well identifying the direction of study comparisons, especially regarding recall. For example, the algorithm failed to identify the comparison direction in a study that reported aminepine to be more effective than clomipramine. The main reason for this low recall is that SemRep’s algorithm to extract directionality is limited to noun phrases in comparative constructions.[9] Specialized NLP techniques that aim at extracting comparison directionality may be needed to improve this performance. Nevertheless, identifying comparative studies and their interventions is the primary step. Defining the directionality is more challenging and not as important as identifying comparative studies and their interventions. In addition, this task could be accomplished by the end user through a knowledge summary user interface. A potential summarization approach is to provide clinicians with a visual representation of comparative studies grouped according to interventions. Table 3 provides a sample output of such a summarization approach.

Limitations

This study has two main limitations. First, the algorithm was evaluated in one case study, therefore it is unknown whether the algorithm would achieve similar performance on other conditions. Second, our algorithm is focused on treatment and cannot be directly used in other types of comparative studies, such as those comparing diagnostic interventions. However, the overall conceptual approach could be extended to other information needs.

Future studies

Areas that warrant future investigation include: 1) fine-tuning and rigorously comparing different machine learning techniques to identify comparative studies; 2) improving intervention extraction performance by integrating information from clinical registries; 3) exploring MeSH data with targeted NLP algorithms to identify combination therapies and different dose/form interventions; 4) exploring targeted NLP methods to extract comparison directions; and 5) assess the algorithm with multiple conditions to test its generalizability.

Conclusion

The algorithm developed in this study achieved a good performance identifying comparative studies from Medline citations and extracting the interventions from these citations. The algorithm did not perform well identifying the direction of intervention comparisons. Further studies are needed to improve intervention extraction and comparison direction. Overall, the proposed method provides the basis for automatic summarization of comparative effectiveness research to support clinicians' decision-making at the point of care.

Acknowledgments

Del Fiol was supported by Agency for Healthcare Research and Quality grant K01HS018352. Jonnalagadda was supported by National Library of Medicine grant 1K99LM011389. The research was also supported by the Intramural Research Program of the NIH, National Library of Medicine.

References

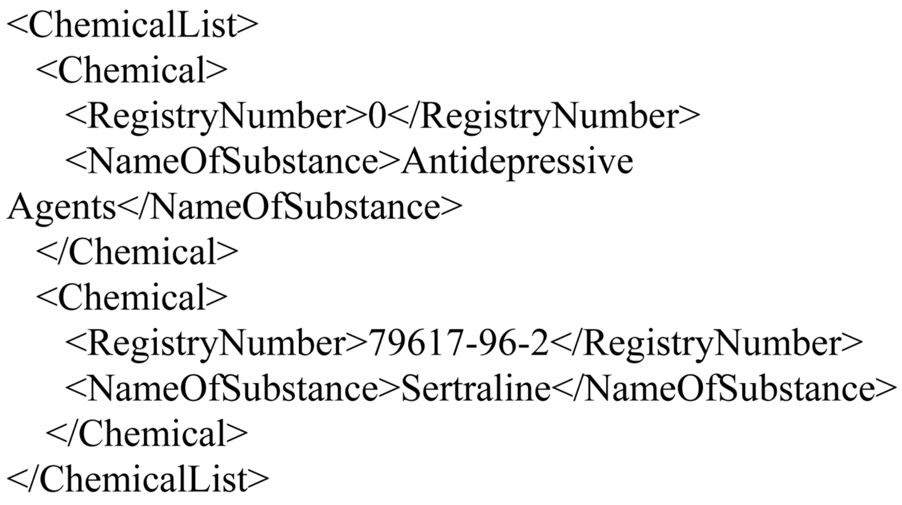
1. Del Fiol G, Workman TE, Gorman PN. Clinicians' Patient Care Information Needs: Preliminary Results of a Systematic Review of the Literature. Proc AMIA Symp. 2012:1605.
2. Ely JW, Osheroff JA, Gorman PN, Ebell MH, Chambliss ML, Pifer EA, Stavri PZ. A taxonomy of generic clinical questions: classification study. BMJ. 2000; 321:429–32. [PubMed: 10938054]
3. Schumock GT, Pickard AS. Comparative effectiveness research: Relevance and applications to pharmacy. Am J Health Syst Pharm. 15;66(14):1278–86.
4. Demner-Fushman D, Hauser SE, Humphrey SM, Ford GM, Jacobs JL, Thoma GR. MEDLINE as a source of just-in-time answers to clinical questions. Proc AMIA Symp. 2006:190–4. [PubMed: 17238329]
5. Jonnalagadda SR, Del Fiol G, Medlin R, Weir C, Fiszman M, Mostafa J, Liu H. Automatically extracting sentences from Medline citations to support clinicians' information needs. J Am Med Inform Assoc. 2012 Epub ahead of print.
6. Kilicoglu H, Shin D, Fiszman M, Rosemblat G, Rindflesch TC. SemMedDB: a PubMed-scale repository of biomedical semantic predications. Bioinformatics. 2012; 28(23):3158–60. [PubMed: 23044550]
7. Lin J, Demner-Fushman D. Semantic clustering of answers to clinical questions. Proc AMIA Symp. 2007:458–62. [PubMed: 18693878]
8. Fiszman M, Demner-Fushman D, Lang FM, Goetz P, Rindflesch TC. Interpreting comparative constructions in biomedical text. Proceedings of the Workshop on BioNLP: Biological, Translational, and Clinical Language Processing. 2007:137–44.

9. Fiszman M, Rindflesch T, Kilicoglu H. Abstraction summarization for managing the biomedical research literature. HLT-NAACL 2004: computational lexical semantic workshop. 2004:76–83.

10. Medline data element descriptions. http://www.nlm.nih.gov/bsd/mms/medlineelements.html

11. Holmes, G.; Donkin, A.; Witten, IH. Weka: a machine learning workbench. Proc Second Australia and New Zealand Conference on Intelligent Information Systems; 1994; Brisbane, Australia.

12. Chung GYC. Towards identifying intervention arms in randomized controlled trials: Extracting coordinating constructions. J Biomed Inform. 2009; 42:790–800. [PubMed: 19166975]



```
<ChemicalList>
  <Chemical>
    <RegistryNumber>0</RegistryNumber>
    <NameOfSubstance>Antidepressive
Agents</NameOfSubstance>
  </Chemical>
  <Chemical>
    <RegistryNumber>79617-96-2</RegistryNumber>
    <NameOfSubstance>Sertraline</NameOfSubstance>
   </Chemical>
</ChemicalList>
```

Figure 1.

Example of a Medline metadata fragment with the RN and NM fields

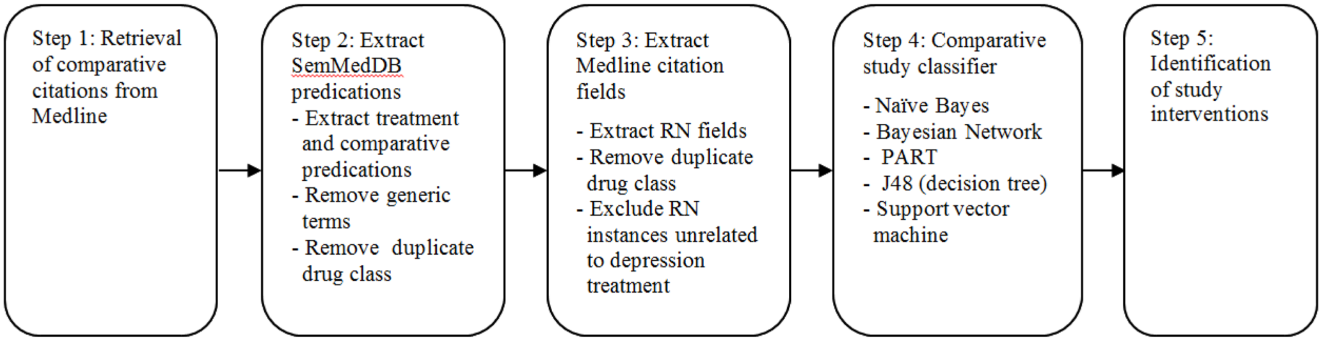

Step 1: Retrieval of comparative citations from Medline
Step 2: Extract SemMedDB predications
- Extract treatment and comparative predications
- Remove generic terms
- Remove duplicate drug class
Step 3: Extract Medline citation fields
- Extract RN fields
- Remove duplicate drug class
- Exclude RN instances unrelated to depression treatment
Step 4: Comparative study classifier
- Naïve Bayes
- Bayesian Network
- PART
- J48 (decision tree)
- Support vector machine
Step 5: Identification of study interventions

Figure 2.

Algorithm Steps

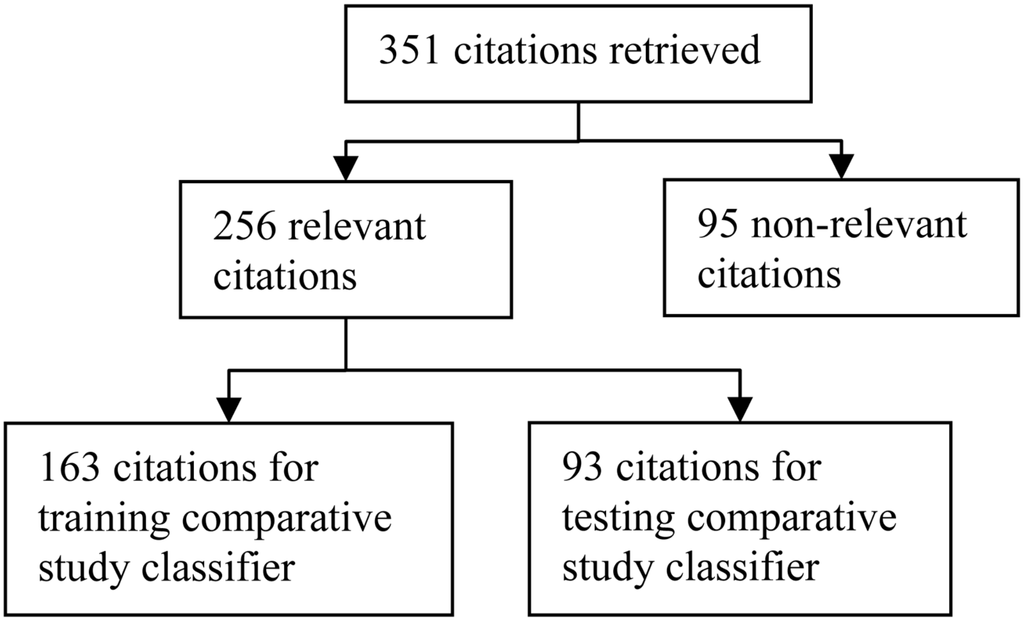

351 citations retrieved
256 relevant citations
95 non-relevant citations
163 citations for training comparative study classifier
93 citations for testing comparative study classifier

Figure 3.
Gold Standard

Table 1

Examples of SemMedDB comparative constructions.

Subject	Predication	Object	Sentence
Mianserin	same_as	Diazepam	In a third trial, mianserin was found to be as effective as diazepam in the treatment of anxiety states in general practice.
Transcranial magnetic stimulation, repetitive	lower_than	Electroconvulsive therapy	RESULTS: Repetitive transcranial magnetic stimulation was significantly less effective than ECT.
Escitalopram	higher_than	Citalopram	Thus, escitalopram is efficacious in depression and the effect occurs earlier than for citalopram.
Fluoxetine	compared_with	Placebos	Depressive symptoms decreased significantly overall with no significant differences between the groups treated with fluoxetine versus placebo.

Table 2

Performance of the comparative study classifiers

	Precision	Recall	F-measure	AUC
Publication Type	0.77	0.58	0.66	N/A
Naïve Bayes	0.83	0.83	0.82	0.90
Bayesian net	0.82	0.82	0.82	0.89
PART	0.83	0.83	0.83	0.76
J48	0.82	0.82	0.82	0.82

Table 3

Sample summarization output

Tricyclic antidepressants • “The data suggest that tryptophan-nicotamide may be as effective as imipramine in unipolar patients providing the dose is kept within the therapeutic window, and that at low doses it could also potentiate the action of tricyclic antidepressants.” • “After 14 days, desipramine prompted an improvement in the Montgomery Asberg Depression Rating Scale (MADRS) score, compared with citalopram and placebo.”
Psychotherapy • “The results suggest that CBT and IPT are robust treatments in both group and individual formats. However, CBT produced significantly greater decreases in depressive symptoms and improved self-concept than IPT.”
rTMS • “The EEG-based interactive technique was associated with an indication of a trend toward a greater clinical effect than the standard rTMS technique. The interactive technique thus has the potential to refine the rTMS methodology and to enhance efficacy in the treatment of depression.”