Proceedings of Medinfo 2010 (in press).

Combining Relevance Assignment with Quality of the Evidence to Support Guideline Development

**Marcelo Fiszman MD PhD,[1] Bruce E. Bray MD,[3] Dongwook Shin PhD,[1] Halil Kilicoglu,[4]
Glen C. Bennett,[2] Olivier Bodenreider MD PhD,[1] Thomas C. Rindflesch PhD[1]**

[1] *National Library of Medicine, NIH, Bethesda, MD*
[2] *National Heart Lung and Blood Institute, NIH, Bethesda, MD*
[3] *Department of Biomedical Informatics, University of Utah, Salt Lake City, UT*
[4] *Department of Computer Science Concordia University, Montréal, QC, Canada*

Abstract

Clinical practice guidelines are used to disseminate best practice to clinicians. Successful guidelines depend on literature that is both relevant to the questions posed and based on high quality research in accordance with evidence-based medicine. Meeting these standards requires extensive manual review. We describe a system that combines symbolic semantic processing with a statistical method for selecting both relevant and high quality studies. We focused on a cardiovascular risk factor guideline, and the overall performance of the system was 56% recall, 91% precision ($F_{0.5}$-score 0.81). If quality of the evidence is not taken into account, performance drops to 62% recall, 79% precision ($F_{0.5}$-score 0.75). We suggest that this system can potentially improve the efficiency of the literature review process in guideline development.

Keywords:

Natural Language Processing, Machine Learning, and Clinical Guidelines

Introduction

Clinical practice guidelines are produced by medical professional societies, governmental agencies, and the biomedical research community to assist clinicians in providing quality care [1, 2, 3]. As part of the guideline creation process, queries are issued to MEDLINE® to retrieve citations relevant to critical questions supporting the guideline. Cited studies must be of high quality (typically, randomized clinical trials and systematic reviews) according to the evidence-based medicine (EBM) paradigm [4]. Currently, domain experts find such studies by reading large numbers of citations for each question posed, a process that is both resource- and time-intensive.

To reduce the amount of manual effort expended during guideline creation, we recently proposed an automatic method based on SemRep [5, 6] semantic processing for discriminating between relevant and nonrelevant MEDLINE citations for critical questions [7]. Although we suggest that the system can help streamline guideline development, it is unable to identify high quality clinical evidence. Others, however, have developed machine learning techniques for automatically recognizing such evidence [8, 9, 10]. In this paper, we describe a system that combines our symbolic processing [7] with a statistical method [10] for selecting studies that are both relevant to questions and of high scientific quality. Retrieved citations are ranked: those which are both relevant and of high quality are put in the highest rank, while those which are nonrelevant and not of high quality are in the lowest. We tested the system on a class of questions for a guideline on cardiovascular risk reduction being produced by the National Heart Lung and Blood Institute at the National Institutes of Health.

Background

SemRep

The SemRep natural language processing system extracts semantic predications from biomedical text using linguistic analysis and domain knowledge in the Unified Medical Language System® (UMLS)® [11]. Processing begins with a partial syntactic analysis based on the SPECIALIST Lexicon [12] and MedPost part-of-speech tagger [13]. Simple noun phrases in this structure are then mapped to Metathesaurus concepts by MetaMap [14]. In final processing, underspecified dependency rules identify some of these augmented noun phrases as arguments of semantic predications asserted in the sentence. Such predications must be sanctioned by a relationship in the UMLS Semantic Network.

For example, from sentence (1), SemRep extracts the predications in (2), in which the arguments (“sibutramine,” “Obesity,” and “Patients”) are concepts from the Metathesaurus, and the relations TREATS and PROCESS_OF are from the Semantic Network.

(1) Second phase of a double-blind study clinical trial on sibutramine for the treatment of patients suffering essential obesity

(2) sibutramine TREATS Obesity
Obesity PROCESS_OF Patients

Predications such as these are used to identify relevant citations for guideline questions.

Relevance

Analysis demonstrated that the distribution of predications representing the semantic components of questions is different in relevant and nonrelevant citations. We therefore wrote rules that match questions to predications [7]. For example, pertinent semantic components of questions in the cardiovascular guideline are risk factor, disorder, population, and action, and the question in (3) has the rules in (4) associated with it.

(3) What is the evidence that diabetes mellitus can be decreased in children?

(4) <Diabetes> PROCESS_OF <Children>
<Diabetes> NOT PROCESS_OF <Adults>
X TREATS <Diabetes>

These rules match SemRep semantic predications in the following way: "X" matches any argument, while items delimited by brackets in the rules represent variables to be matched to a specified domain of Metathesaurus concepts serving as arguments of predications. For example, "<Diabetes>" matches concepts such as "Diabetes Mellitus," "Diabetes Mellitus, Insulin-Dependent," and "Diabetes Mellitus, Non-Insulin-Dependent," while "<Children>" matches "Child," "Youth," "Boys," and "Girls."

If predications matching the rules in (4) are found in the SemRep output for a retrieved citation, it is relevant to the question; otherwise, it is nonrelevant. It is important to note that this processing does not answer the question (e.g. [15-19]), but merely identifies citations which can be used to determine an answer. The word *evidence* in the question implies that only high-quality evidence is sought.

Quality of the Evidence

The evidence-based medicine paradigm categorizes types of clinical evidence and ranks them according to their strength (quality) to avoid research bias. The Oxford Centre for Evidence-based Medicine [20] proposed an influential categorization which states that the most rigorous scientific studies are systematic reviews and randomized clinical trials. The second component of our combined system identifies such studies, using machine-learning techniques trained on MEDLINE citations annotated by hand for high-quality evidence [21, 22]. An array of features, including text words, semantic features, and MEDLINE metadata are used by several classifiers augmented with boosting and ensemble techniques to mark each document encountered as either reporting high quality research or not. The classifier achieved an F_1 score of nearly 0.70 in making this determination in test documents [10].

Materials and Methods

Our technique for combining symbolic relevance processing with the statistical method for identifying high-quality research is tied to a particular guideline question. A PubMed query is issued to retrieve citations for that question. Relevance and quality of evidence processing are then applied independently to the retrieved citations and the results are combined so that each citation has a score both for relevance (1 or 0) and for quality of evidence (also 1 or 0). Citations are ranked into four disjoint groups based on these scores:

A. Relevant to the question and high quality of evidence (1,1)

B. Relevant to the question but not high quality of evidence (1,0)

C. Nonrelevant to the question but high quality of evidence (0,1)

D. Nonrelevant to the question and not high quality of evidence (0,0)

Citations in A have the highest probability of being true positives and those in D the lowest. We hypothesized that we could exploit the A group to retain only systematic reviews and randomized clinical trials in the list of relevant citations recommended to the guideline developers. To test this hypothesis we first calculated recall and precision on the A and B group combined and compared these metrics to those computed on the A group alone.

In constructing a reference standard for evaluation, we selected four questions from the guideline on cardiovascular disease risk reduction and issued a PubMed query for each. All queries were limited in PubMed to: Only items with abstracts, Humans, Meta-Analysis, Randomized Controlled Trial, and English. A further limit was imposed appropriate to the target population of the question. Questions, queries, and additional limits were as follows:

Question 1. What is the evidence for the effect of sibutramine on weight loss and maintenance in adults?
Query: sibutramine
Additional limits: All Adult: 19+ years
Retrieved citations annotated: 91

Question 2. What is the evidence that obesity can be decreased in children?
Query: Obesity/therapy[majr]
Additional limits: All Child: 0-18 years
Retrieved citations annotated: 100

Question 3. What is the evidence that hyperlipidemia can be decreased in children?
Query: Hyperlipidemia/therapy[majr]
Additional limits: All Child: 0-18 years
Retrieved citations annotated: 88

Question 4. What is the evidence that diabetes mellitus can be decreased in children?
Query: Diabetes Mellitus/therapy[majr]
Additional limits: All Child: 0-18 years
Retrieved citations annotated: 100

Retrieved citations retrieved were then annotated by the first and second authors as being relevant (or not) and high quality (or not). Limiting their analysis to titles and abstracts, relevant citations were those considered to be informative in answering the question. System output for Group A output and for combined A and B output were compared to the reference standard separately, and recall, precision, and F_1-score were calculated. We also calculated a weighted $F_{0.5}$-score (1.25*P*R)/(0.25*P +R), which weights precision twice as much as recall, since guideline developers value precision more highly than recall.

Results

For evaluation, a citation had to be marked as both relevant and high quality in order to be considered a true positive, and, of the 379 total citations processed, 138 were so marked in the reference standard. We first consider overall results for all questions for Groups A and B combined. These two groups include all citations returned as relevant by the system, and are equivalent to not exploiting quality of the evidence processing. The system missed 52 (out of 138) of these, yielding recall of 62%. Of the 108 citations returned, 22 were false positives (either nonrelevant or not high quality), resulting in precision of 79%. The F_1 score was 0.69 and the $F_{0.5}$ was 0.75. The $F_{0.5}$ scores for the individual questions in Groups A and B combined range from 0.59 to 0.90. (See Table 1.)

Table 1. Questions and performance metrics for combined Groups A and B, N=Total number of citations. R = Recall, P = Precision

Question	N	R	P	F_1	$F_{0.5}$
Question 1	91	50%	82%	0.63	0.73
Question 2	100	52%	61%	0.56	0.59
Question 3	88	86%	90%	0.88	0.90
Question 4	100	76%	84%	0.80	0.83
Overall	**379**	**62%**	**79%**	**0.69**	**0.75**

As noted, citations in Group A are marked as both relevant and high quality by the system. In system output for Group A, 61 relevant citations were missed, resulting in overall recall of 56%. For all questions, the system marked 85 citations as belonging to Group A; of these, 8 were false positives, in that they were not both relevant and high quality, resulting in precision of 91% and an $F_{0.5}$ score of 0.81 for this group (an improvement of 6% over results for Groups A and B combined). The $F_{0.5}$ scores in Group A range from 0.67 to 0.92 for the individual questions. Although precision was higher in Group A alone than in A and B combined, lower recall caused the F_1 score to remain constant. (See Table 2.)

Table 2. Questions and performance metrics for Group A. N=Total number of citations. R = Recall, P = Precision

Question	N	R	P	F_1	$F_{0.5}$
Question 1	91	45%	96%	0.62	0.78
Question 2	100	42%	78%	0.55	0.67
Question 3	88	82%	95%	0.88	0.92
Question 4	100	72%	91%	0.81	0.87
Overall	**379**	**56%**	**91%**	**0.69**	**0.81**

Discussion

As shown in Tables 1 and 2, recall was always lower than precision and this is mostly due to SemRep missing predications. However, there were also mistakes made by the machine learning system when assigning high quality citations. Since it is hard to determine the real cause of machine learning errors, we give a SemRep example. (6) was the only predication generated for (5), which occurred in a relevant citation.

(5) This study demonstrates that plant stanols reduce LDL-C levels in children with hypercholesterolemia

(6) Hypercholesterolemia PROCESS_OF Child

The citation in which (5) occurred was not marked as relevant, primarily because "plant stanols" does not occur in the Metathesaurus, and thus SemRep could not extract the predication for stanols reducing LDL-C levels. Moreover, an inference based on domain knowledge would be required to determine that reducing LDL-C levels is related to decreasing hypercholesterolemia (hence making this citation relevant), and this is beyond the current capability of SemRep.

Our preliminary evaluation suggests that combining relevance assignment with quality of the evidence processing has promise in supporting guideline development. Although recall in the combined A and B group was 6% higher than in Group A alone, precision in Group A was 91% as compared to 79% for group A and B combined. The $F_{0.5}$-score rose from 0.75 (A and B) to 0.81 (group A). Guideline developers strive for high precision because of the added expense of reading additional citations. Although preliminary, our findings support the claim that considering citations only in group A sacrifices very few citations that are both relevant and based on high-quality studies. Additional strategies exploiting groups A, B, C, D might help guideline developers in other ways. The ranking inherent in this partitioning could be used to prioritize reading and conserve resources.

Group D (nonrelevant and not high quality), the lowest rank, may be useful for reducing the number of citations to be read during guideline development. Considering all questions, the system assigned 56 (out of 379) citations to this Group. Eleven of these were both relevant and high quality, as opposed to 77 out of 85 in the A group. It is unlikely to be worth developer time to read citations in Group D.

Group C presents a trade-off between sacrificing true positives and effort spent. The system assigned 215 citations to this (nonrelevant but high quality) group. Of these, 41 were true positives according to the reference standard. This is a considerably high number; however, in order to find the true positives, developers would have to read 215 citations, which may not be cost effective.

Another consideration is that for some questions, developers are interested in case-controls, observational, and even serialized case report studies. For such questions, an appropriate partition would be A and B for true positives and C and D for true negatives.

This study is limited in that the evaluation was based on a small sample of four questions and was not conducted in the context of actual guideline development. Further, the method was tested on only one class of question and it remains to be seen how the incorporation of quality of the evidence processing will extend to other question classes.

Conclusion

In the context of clinical practice guideline development, we describe a system that combines symbolic semantic processing with a statistical method for selecting studies that are both relevant to guideline questions and of high scientific quality, the most valuable research for guideline developers. Evaluation revealed that exploiting the combined processing allowed us to improve overall performance by 6%. Finally, we described how this combined system might be useful to support guideline development.

Acknowledgments

This study was supported in part through an interagency agreement between the National Library of Medicine and the National Heart Lung and Blood Institute and by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

References

[1] Institute of Medicine. (1990). Clinical Practice Guidelines: Directions for a New Program. M.J. Field and K.N. Lohr (eds.) Washington, DC: National Academy Press.

[2] Institute of Medicine, Committee on Quality Health Care in America. (2001). Crossing the quality chasm: A new health system for the 21st century. Washington, DC: National Academy Press.

[3] http://www.nhlbi.nih.gov/guidelines/index.htm

[4] Sackett DL, Straus SE, Richardson WS, et al. Evidence-Based Medicine: How to Practice and Teach EBM. Philadelphia, PA, Churchill Livingstone, 2000

[5] Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J of Biomed Inf. 2003 Dec;36(6):462-77.

[6] Rindflesch TC, Fiszman M, Libbus B. Semantic interpretation for the biomedical research literature. In Chen, Fuller, Hersh, and Friedman, eds. Medical informatics: Knowledge management and data mining in biomedicine. Springer, 2005, pp. 399-422

[7] Fiszman M, Ortiz E, Bray BE, Rindflesch TC. Semantic processing to support clinical guideline development. AMIA Annu Symp Proc. 2008 Nov 6:187-91.

[8] 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; 2: 207-16.

[9] Cohen A. Optimizing feature representation for automated systematic review work prioritization. AMIA Annu Symp Proc. 2008 Nov 6:121-5.

[10] Kilicoglu H, Demner-Fushman D, Rindflesch TC, Wilczynski NL, Haynes RB. Towards automatic recognition of scientifically rigorous clinical research evidence. J Am Med Inform Assoc. 2009 Jan-Feb;16(1):25-31.

[11] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D267-70.

[12] McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. Proc Annu Symp Comput Appl Med Care 1994;235-9.

[13] Smith L, Rindflesch T, Wilbur WJ. MedPost: a part-of-speech tagger for biomedical text. Bioinformatics 2004;20(14):2320-1.

[14] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. Proc AMIA Symp. 2001;17-21.

[15] Demner-Fushman D, Lin J. Answering clinical questions with knowledge-based and statistical techniques. Computational Linguistics. 2007 Mar; 33 (1):63-103.

[16] Niu Y, Zhu X, Hirst G. Using outcome polarity in sentence extraction for medical question-answering. AMIA Annu Symp Proc. 2006;599-603.

[17] Yu H, Lee M, Kaufman D, et al. Development, implementation, and a cognitive evaluation of a definitional

question answering system for physicians. J Biomed Inform. 2007 Jun;40(3):236-51.

[18] Huang X, Lin J, Demner-Fushman D Evaluation of PICO as a knowledge representation for clinical questions. AMIA Annu Symp Proc. 2006;359-63

[19] 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 Jun 15;7:16.

[20] http://www.cebm.net/index.aspx?o=1025#levels

[21] Wong SS, Wilczynski NL, Haynes RB. Developing optimal search strategies for detecting clinically relevant qualitative studies in MEDLINE. Medinfo 2004; 11:311-6

[22] 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 Nov-Dec;1 (6):447-58.

Address for correspondence

Marcelo Fiszman, M.D. Ph.D. National Library of Medicine 8600 Rockville Pike, Bldg 38A, Rm B1N-28J, Bethesda, MD 20894, Email: fiszmanm@mail.nih.gov