Evidence-Based Medicine, the Essential Role of Systematic Reviews, and the Need for Automated Text Mining Tools

Aaron M. Cohen[1*], Clive E. Adams[2], John M. Davis[3], Clement Yu[4], Philip S. Yu[4], Weiyi Meng[5], Lorna Duggan[6], Marian McDonagh[1], and Neil R. Smalheiser[3]

[1] Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, Oregon, USA.

[2] Mental Health Services Research & Cochrane Schizophrenia Group, University of Nottingham, UK.

[3] Department of Psychiatry, UIC Psychiatric Institute MC912, University of Illinois at Chicago, Chicago, IL, USA.

[4] Department of Computer Science, University of Illinois at Chicago, Chicago, Illinois, USA.

[5] Department of Computer Science, Binghamton University, Binghamton, NY, USA.

[6] Care Principles Ltd, Oaks Lodge, Fordham Road, Newmarket, Suffolk, UK.

ABSTRACT

High quality, cost-effective medical care requires consideration of the best available, most appropriate evidence in the care of each patient, a practice known as Evidence-based Medicine (EBM). EBM is dependent upon the wide availability and coverage of accurate, objective syntheses called evidence reports (also called systematic reviews). These are compiled by a time and resource-intensive process that is largely manual, and that has not taken advantage of many of the advances in information processing technologies that have assisted other textual domains. We propose a specific text-mining based pipeline to support the creation and updating of evidence reports that provides support for the literature collection, collation, and triage steps of the systematic review process. The pipeline includes a metasearch engine that covers both bibliographic databases and selected "grey" literature; a module that classifies articles according to study type; a module for grouping studies that are closely related (e.g. that derive from the same underlying clinical trial or same study cohort); and an automated system that ranks publications according to the likelihood that they will meet inclusion criteria for the report. The proposed pipeline will also enable groups performing systematic review to reuse tools and models created by other groups, and will provide a test-bed for further informatics research to develop improved tools in the future. Ultimately, this should increase the rate that high-quality systematic reviews and meta-analyses can be generated, accessed and utilized by clinicians, patients, care-givers, and policymakers, resulting in better and more cost-effective care.

IHI'10, November 11–12, 2010, Arlington, Virginia, USA.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering, Search process, Selection process

J.3 [**Life and Medical Sciences**]: Medical information systems

General Terms

Algorithms, Design, Management

Keywords

Information Storage and Retrieval, Text-Mining, Evidence-Based Medicine.

1. Introduction

The practice of Evidence-Based Medicine (EBM) involves judicious use of the best and most up-to-date evidence, in the form of published literature, to patient care decision making (Sackett, Rosenberg, Gray, Haynes, & Richardson, 1996). While the original vision of EBM appeared to require physicians to search the primary literature for evidence applicable to their patients, the modern conception of EBM relies heavily on distillations or summaries of the primary literature in the form of systematic reviews (also called evidence reports) (A.M. Cohen, Stavri, & Hersh, 2004; Hersh, 1999). Both medical practitioners and policy makers regularly utilize these summaries when making important decisions (Helfand, 2005a, 2005b). There are now many collections of these reviews including the Cochrane Collaboration (e.g., Adams, et al., 2008) and the Evidence-based Practice Centers (EPCs) of the Agency for Healthcare Research and Quality (AHRQ) (Grimshaw, Santesso, Cumpston, Mayhew, & McGowan, 2006; Helfand, 2005a). The creation of these evidence reports requires the efforts of experts known as systematic reviewers, and is a time consuming, demanding and largely manual resource-intensive process. Because the progress of medical science is ever-expanding, there is an ongoing need both for more evidence reports and for periodic updating of existing reports. In the near future, when genome-based personalized medicine becomes standard practice, it will be even

more important to identify and summarize the best available evidence specific to each person and their individual situation. Here, we propose a multi-step text mining based pipeline to facilitate monitoring, production, and maintenance of summaries of the best available medical evidence for a wide range of medical conditions.

2. Scope of the Problem

Writing an evidence report is a labor-intensive process requiring weeks to months of human effort. In any systematic review, potentially thousands of articles must be located, triaged, reviewed, and summarized. Potentially relevant articles are located using an iteratively refined search of biomedical electronic databases, such as MEDLINE and EMBASE. Articles are then triaged in a two-step process. First, the abstract is reviewed and, if the abstract suggests that the full paper should be inspected, the entire article is then read. Second, if the full text article proves to meet the inclusion criteria, the evidence presented in the article is summarized and included in the systematic review.

As can be seen from the graph in Figure 1, reports of new clinical trials are being published at the rate of over 20,000 per year. In recent years, about 3000 systematic reviews and evidence reports and updates were indexed by the National Library of Medicine for MEDLINE per year. The Cochrane Collaboration has estimated that at least 10,000 separate and up-to-date systematic reviews are needed to cover most common health care problems. Currently, less than half of these have been completed (Mallett & Clarke, 2003). The challenge posed to the systematic review community is already large and continuing to grow.

Evidence reports are most useful when their conclusions are based on the most up to date research (Atkins, Fink, & Slutsky, 2005). Updating a systematic review may require as many resources as the creation of a new review. There is, however, little agreement on how or when to update a review (Moher, et al., 2007, 2008; Shekelle, Eccles, Grimshaw, & Woolf, 2001). A recent systematic review of literature studying when and how to update systematic reviews found little published research (Moher, et al., 2007). The authors argued that: "More research is needed to develop pragmatic and efficient methodologies for updating systematic reviews". It is clear that new tools are needed to help identify and process new relevant evidence in a timely manner.

Perhaps surprisingly, the workflow of most systematic review groups is largely a manual process, with much duplication of effort, and without sharing systems and solutions between groups. Search strategies are created from scratch for each topic.

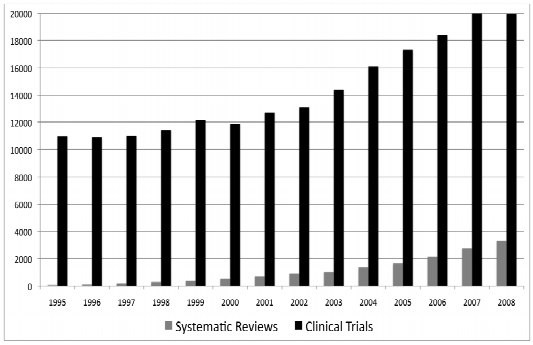

Figure 1. Number of reports on clinical trials and systematic reviews indexed in MEDLINE by year.

Although search engines may be used to search online databases, generally each database is searched independently, and there is little automated help with the inevitable duplication of hits or the identification of multiple reports that are based on the same underlying data. No highly refined automated tools are available to help reviewers sort and prioritize articles. The contribution of each and every article returned by the initial search is determined only after each of the team reviewers read and summarize the abstract and possibly the full text. Inclusion/exclusion decisions are tracked in reference manager software but are not generally used to refine the process when the review is updated. Different groups conducting reviews in related areas lack shared resources. It has been estimated that at least 350 of the ~ 3,500 hours spent on a systematic review consists of reviewing the references returned from the initial literature search (A.M. Cohen, Hersh, Peterson, & Yen, 2006). Improving this process by even 10% can result in saving a person-week worth of work for each review. This time can be productively spent performing meta-analyses, improving the final report, designing the definitive relevant evaluative study, or starting work on another review.

Whereas other areas of annotation and information summarization, such as genome annotation, are very active topics of research in the bioinformatics community (Hersh, et al., 2006; Hirschman, Yeh, Blaschke, & Valencia, 2005; Rzhetsky, Shatkay, & Wilbur, 2009), there is relatively little published research on building automated tools to assist the systematic review process. However, there is a solid foundation of research in the areas of information retrieval and automated document classification to support the practice of EBM and much of this work is applicable to the systematic review process. Haynes et al. created the PubMed Clinical Queries, template-based Boolean queries that support clinicians practicing EBM; the group has continued to refine additional search strategies to retrieve high quality causation and diagnostic studies (R. B. Haynes, Wilczynski, McKibbon, Walker, & Sinclair, 1994; Wilczynski & Haynes, 2003, 2005; Wilczynski, Morgan, & Haynes, 2005). Kilicoglu et al. (Kilicoglu, Demner-Fushman, Rindflesch, Wilczynski, & Haynes, 2008, 2009) and Aphinyanaphongs et al. (Aphinyanaphongs & Aliferis, 2003; Aphinyanaphongs, Statnikov, & Aliferis, 2006) have trained machine learning systems to identify studies that are high quality according to EBM standards of evidence. Cohen et al. have created topic specific machine learning models that accurately predict inclusion of articles in specific systematic reviews, and have shown that the performance of topic specific models is higher than general models of systematic review inclusion (A.M. Cohen, 2008; A. M. Cohen, Ambert, & McDonagh, 2009; A.M. Cohen, et al., 2006).

3. A Text-Mining Pipeline for EBM

Figure 2 (below) presents a high level architectural view of the proposed text-mining pipeline for EBM. The main functionalities are represented by the four arrows labeled *Search*, *Classify*, *Group*, and *Rank*.

The **first component** of the pipeline (labeled *Search)* is the **metasearch engine**. This is the sub-system that accepts topic specific queries, routes them to multiple search engines, collects the individual results and produces an integrated list of matched results. The external databases shown are interfaced to the metasearch engine in order to allow a single query input to be simultaneously distributed to multiple on-line searchable

databases, and the results collected for further pipeline processing. Our own group has identified PubMed, EMBASE, PsycINFO, CINAHL, and the Cochrane Central Register of Controlled Trials as a core set of key databases, but other databases can be included in the same manner as the need arises.

Other information sources, such as important government web sites and other sources of gray literature, are more difficult to identify in online search engines because they reside in scattered sites that may not have permanent URLs, and because high-quality documents are difficult to identify automatically in the face of a large excess of "noise". We suggest that collections of non-published high-quality documents may be identified manually, indexed using a consistent set of metadata, and placed into an internal database that is then included in the set of databases accessed by the metasearch engine. The engine will also detect duplicate articles from multiple sources.

This approach builds on a substantial amount of successful metasearch engine work published both in the information retrieval literature (e.g., (Liu, et al., 2007; Lu, Meng, Shu, Yu, & Liu, 2005)) as well as the medical informatics literature (e.g., (Coiera, Westbrook, & Rogers, 2008)). However, some significant research challenges will need to be solved in order to ensure that the metasearch engine has very high recall while maintaining reasonable precision. For example, the system will need to provide de-duplication of search results. The system will need to support a variety of types of queries, e.g., time sliced data or restriction to human studies. Furthermore, it will need to carry out query processing to optimize the way that terms are handled that correspond to diseases, drugs, authors, geographical names, abbreviations, synonyms, etc. Moreover, the queries need to be adjusted according to the query model for each search engine. At present, we are focusing on a core set of 5 relatively stable search engines that can be connected manually, but it would be desirable to develop mechanisms that can automatically connect to and adapt to new search engines.

The **second component** in the pipeline is the **Classifier** (labeled *Classify*). This will label articles in terms of their type: randomized controlled trial, review article, evidence report, clinical guideline, or other, corresponding to the Haynes "5S" evolution of information service type (B. Haynes, 2006). Support vector machine (SVM) based classification systems have been shown to perform well on sparse feature-space text classification tasks. Furthermore, extensions to SVM, such as one-against-the-rest and error correcting output codes (Dietterich & Bakiri, 1995) provide an effective means of using SVM on multi-way classification tasks such as this (e.g., (A. M. Cohen, 2008)). Furthermore, while standard SVM formulations do not allow incremental training, in this application the models can be trained and retrained in batches as portions of review updates are completed. Additionally, both incremental techniques and those based on restricting the prior data to the support vectors, have been successfully applied in other domains and can be used to make retraining much more efficient in this application as well (Domeniconi & Gunopulos, 2001; Laskov, Gehl, Krueger, & Mueller, 2006), allowing the training and application of these models on a large scale.

The **third component** in the pipeline is the **Aggregator** (labeled *Group*). This will group (or cluster) articles in terms of publication relatedness. Publications may be related because they represent duplicate publications, arise from the same study protocol, or use the same study cohort. This problem, while currently under-explored, has similar qualities to the name-disambiguation challenge of clustering Medline articles written by the same individual and can likely be approached in a similar manner (Torvik & Smalheiser, 2009). Briefly, articles written by the same person tend to share certain informative characteristics that suffice to distinguish them from articles written by other people with the same name. To create gold standard sets of related articles for training the Aggregator, manually clustered sets of articles as well as sets of articles that share the same Clinical Trial Registry numbers can be employed.

The **fourth component** in the pipeline is the **Prioritizer** (labeled *Rank*). This component ranks the classified and grouped articles forwarded by the prior pipeline stages according to their likelihood for inclusion in the systematic review. This ranking uses current information within those articles, as well as a model built using the past history of including other articles. The past history that the predictions are based on can be customized for an individual reviewer, review team, or topic, and can also use a general model based on all of the review inclusion data available to the system. Furthermore, the publication type of the article can be taken into account, as well as related articles determined by the grouping function. In this way, the ranking models can be productively used by any potential user, although the models will perform better when customized based on past review work relevant to that user, review team, and systematic review topic (A. M. Cohen, et al., 2009). The very closely related prior cited work of Cohen et al., Kilicoglu et al., and Aphinyanaphongs et al. all provide strong evidence that this task can be successfully accomplished with SVM-based classifiers using a combination of text and metadata based (e.g., MeSH) features.

The output of the pipeline will be shown to the user in an interface specifically designed in collaboration with systematic review experts. The interface will allow annotation and the import and export of search, ranking, and annotation information to other tools that reviewers may use, such as EndNote (www.endnote.com) or other bibliographic managers. The user interface will also provide a key function for system improvement by keeping track of inclusion/exclusion decisions along with other annotations, and by collecting usage information from the users. Thus, it will be possible to improve both the machine learning models and the usability and usefulness of the system as a whole.

4. Future Work and Long-term Vision

Once constructed the usability and utility of the pipeline approach will be evaluated and iteratively refined. The usability can be evaluated using established techniques such as discount usability testing. The effectiveness of the system can be evaluated using both time and effort measurements (comparing to previous similar reviews and updates performed without the pipeline), as well as user satisfaction surveys.

Each review group may have their own detailed workflow, but to a great extent, the first steps in conducting a systematic review, such as identifying and filtering literature for inclusion, are consistent across groups and independent of the review topic. An integrated pipeline can benefit all systematic review groups by gaining economies and amortizing the overhead in data collection

and model building. Using these models the pipeline may also be useful for other EBM-related activities with literature triage steps such as clinical guideline development.

Ideally, the pipeline would be implemented as a public utility that can be accessed via a universal web-based interface. However, because different users have different access to subscription-based sources of information, it may be necessary to create stand-alone software that each user group can implement at their own location (though user-created decisions and annotations can still be sent to a single home site to allow the system to function and evolve as a whole).

We view the pipeline not simply as a practical tool, but as a test-bed for conducting further fundamental research into informatics issues. For example, conducting cross-language retrieval of non-English articles lies beyond the current state of the art, but will be an important future goal as clinical trials in non-English speaking countries (e.g., China) become progressively more visible. Similarly, the proposed system does not specifically address the issue of automatically extracting high quality information present within summaries or full-text of individual articles (Demner-Fushman, Few, Hauser, & Thoma, 2006; Mendonca & Cimino, 2000), nor does it include automated methods of synthesizing and weighing (possibly conflicting) evidence across a set of articles. However, having an active pipeline (with ongoing annotation of reviewer decisions) will facilitate such studies. Lastly, the pipeline system will provide a means of studying and addressing the question of when best to update a systematic review.

In conclusion, the text mining-based pipeline for accelerating systematic reviews in evidence-based medicine will decrease the manual burden of systematic reviewers during the literature collection and review process, and increase the proportion of reviewer time spent synthesizing evidence, performing meta-analyses, and considering results. It should help to improve the coverage, dissemination, and acceptance of evidence-based medicine within the biomedical community. Ultimately, this should lead to better and more cost-effective health care.

5. REFERENCES

Adams, C. E., Coutinho, E. S., Davis, J., Duggan, L., Leucht, S., Li, C., et al. (2008). Cochrane Schizophrenia Group. *Schizophr Bull, 34*(2), 259-265.

Aphinyanaphongs, Y., & Aliferis, C. F. (2003). Text categorization models for retrieval of high quality articles in internal medicine. *AMIA Annu Symp Proc*, 31-35.

Aphinyanaphongs, Y., Statnikov, A., & Aliferis, C. F. (2006). A comparison of citation metrics to machine learning filters for the identification of high quality MEDLINE documents. *J Am Med Inform Assoc, 13*(4), 446-455.

Atkins, D., Fink, K., & Slutsky, J. (2005). Better information for better health care: the Evidence-based Practice Center program and the Agency for Healthcare Research and Quality. *Ann Intern Med, 142*(12 Pt 2), 1035-1041.

Cohen, A. M. (2008). Five-way Smoking Status Classification using Text Hot-spot Identification and Error-Correcting Output Codes. *J Am Med Inform Assoc, 15*(1), 32-35.

Cohen, A. M. (2008). Optimizing feature representation for automated systematic review work prioritization. *AMIA Annu Symp Proc*, 121-125.

Cohen, A. M., Ambert, K., & McDonagh, M. (2009). Cross-topic Learning for Work Prioritization in Systematic Review Creation and Update. *J Am Med Inform Assoc, 16*(5), 690-704.

Cohen, A. M., Hersh, W. R., Peterson, K., & Yen, P. Y. (2006). Reducing workload in systematic review preparation using automated citation classification. *J Am Med Inform Assoc, 13*(2), 206-219.

Cohen, A. M., Stavri, P. Z., & Hersh, W. R. (2004). A Categorization and Analysis of the Criticisms of Evidence-Based Medicine. *Int J Med Inf, 73*(1), 35-43.

Coiera, E., Westbrook, J. I., & Rogers, K. (2008). Clinical decision velocity is increased when meta-search filters enhance an evidence retrieval system. *J Am Med Inform Assoc, 15*(5), 638-646.

Demner-Fushman, D., Few, B., Hauser, S. E., & Thoma, G. (2006). Automatically identifying health outcome information in MEDLINE records. *J Am Med Inform Assoc, 13*(1), 52-60.

Dietterich, T. G., & Bakiri, G. (1995). Solving Multiclass Learning Problems via Error-Correcting Output Codes. *Journal of Artificial Intelligence Research*, 263-286.

Domeniconi, C., & Gunopulos, D. (2001). *Incremental support vector machine construction.* Paper presented at the Proceedings of the 1st international conference on data mining (ICDM), San Jose, CA.

Grimshaw, J. M., Santesso, N., Cumpston, M., Mayhew, A., & McGowan, J. (2006). Knowledge for knowledge translation: the role of the Cochrane Collaboration. *J Contin Educ Health Prof, 26*(1), 55-62.

Haynes, B. (2006). Of studies, syntheses, synopses, summaries, and systems: the "5S" evolution of information services for evidence-based healthcare decisions. *British Medical Journal, 11*(6), 162.

Haynes, R. B., Wilczynski, N., McKibbon, K. A., Walker, C. J., & Sinclair, J. C. (1994). Developing optimal search strategies for detecting clinically sound studies in MEDLINE. *J Am Med Inform Assoc, 1*(6), 447-458.

Helfand, M. (2005a). Incorporating information about cost-effectiveness into evidence-based decision-making: the evidence-based practice center (EPC) model. *Med Care, 43*(7 Suppl), 33-43.

Helfand, M. (2005b). Using Evidence Reports: Progress And Challenges In Evidence-Based Decision Making. *Health Affairs, 24*(1), 123-127.

Hersh, W. (1999). "A world of knowledge at your fingertips": the promise, reality, and future directions of on-line information retrieval. *Acad Med, 74*(3), 240-243.

Hersh, W., Bhupatiraju, R. T., Ross, L., Roberts, P., Cohen, A. M., & Kraemer, D. F. (2006). Enhancing access to the bibliome: the TREC 2004 Genomics Track. *Journal of Biomedical Discovery and Collaboration, 1*(3).

Hirschman, L., Yeh, A., Blaschke, C., & Valencia, A. (2005). Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics, 6*(Suppl 1), S1.

Kilicoglu, H., Demner-Fushman, D., Rindflesch, T. C., Wilczynski, N. L., & Haynes, R. B. (2008). Toward automatic recognition of high quality clinical evidence. *AMIA Annu Symp Proc*, 368.

Kilicoglu, H., Demner-Fushman, D., Rindflesch, T. C., Wilczynski, N. L., & Haynes, R. B. (2009). Towards automatic recognition of scientifically rigorous clinical research evidence. *J Am Med Inform Assoc, 16*(1), 25-31.

Laskov, P., Gehl, C., Krueger, S., & Mueller, K. (2006). Incremental support vector learning: Analysis, implementation and applications. *The Journal of Machine Learning Research, 7*, 1936.

Liu, K., Meng, W., Qui, J., Yu, C., Raghavan, V., Wu, Z., et al. (2007). *AllInOneNews: Development and Evaluation of a Large-Scale News Metasearch Engine.* Paper presented at the ACM SIGMOD International Conference on Management of Data ACM (SIGMOD 2007), Industrial track, Beijing, China.

Lu, Y., Meng, W., Shu, L., Yu, C., & Liu, K. (2005). *Evaluation of result merging strategies for metasearch engines.* Paper presented at the 6th International Conference on Web Information Systems Engineering (WISE05).

Mallett, S., & Clarke, M. (2003). How many Cochrane reviews are needed to cover existing evidence on the effects of health care interventions? *ACP J Club, 139*(1), A11.

Mendonca, E. A., & Cimino, J. J. (2000). Automated knowledge extraction from MEDLINE citations. *Proc AMIA Symp*, 575-579.

Moher, D., Tsertsvadze, A., Tricco, A. C., Eccles, M., Grimshaw, J., Sampson, M., et al. (2007). A systematic review identified few methods and strategies describing when and how to update systematic reviews. *Journal of Clinical Epidemiology, 60*(11), 1095-1095.

Moher, D., Tsertsvadze, A., Tricco, A. C., Eccles, M., Grimshaw, J., Sampson, M., et al. (2008). When and how to update systematic reviews. *Cochrane Database Syst Rev*(1), MR000023.

Rzhetsky, A., Shatkay, H., & Wilbur, W. J. (2009). How to Get the Most out of Your Curation Effort. *PLoS Computational Biology, 5*(5), e1000391.

Sackett, D. L., Rosenberg, W. M., Gray, J. A., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn't. *BMJ, 312*(7023), 71-72.

Shekelle, P., Eccles, M. P., Grimshaw, J. M., & Woolf, S. H. (2001). When should clinical guidelines be updated? *BMJ, 323*(7305), 155-157.

Torvik, V. I., & Smalheiser, N. R. (2009). Author Name Disambiguation in MEDLINE. *ACM Trans Knowl Discov Data, 3*(3):11.

Wilczynski, N. L., & Haynes, R. B. (2003). Developing optimal search strategies for detecting clinically sound causation studies in MEDLINE. *AMIA Annu Symp Proc*, 719-723.

Wilczynski, N. L., & Haynes, R. B. (2005). EMBASE search strategies for identifying methodologically sound diagnostic studies for use by clinicians and researchers. *BMC Medicine, 3*(1), 7.

Wilczynski, N. L., Morgan, D., & Haynes, R. B. (2005). An overview of the design and methods for retrieving high-quality studies for clinical care. *BMC Med Inform Decis Mak, 5*, 20.

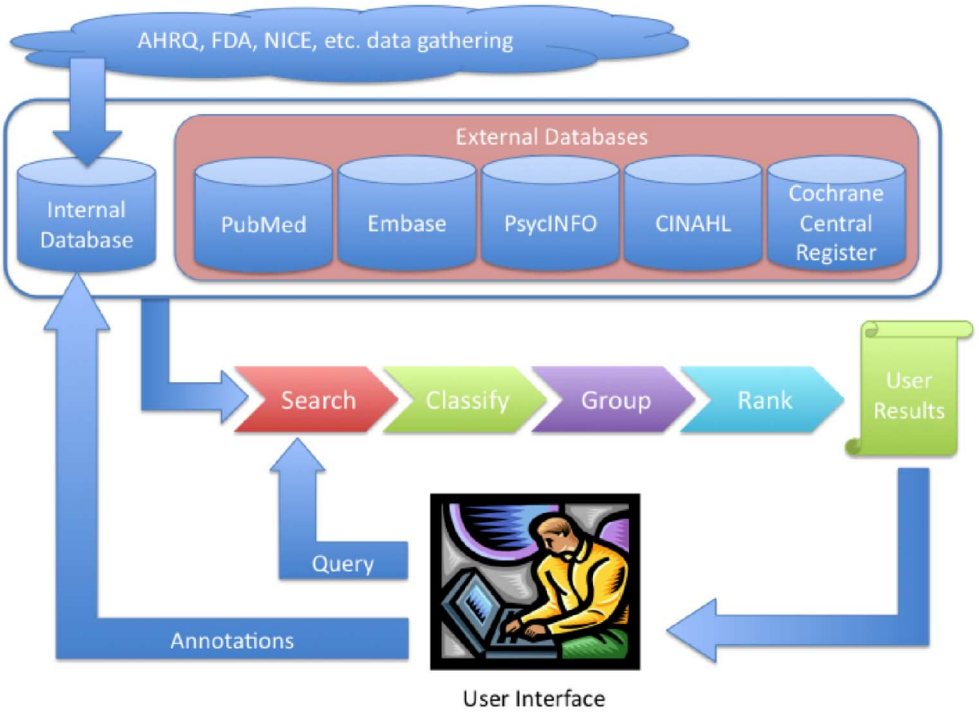

Figure 2. Text Mining Pipeline Architecture to Support Evidence-Based Systematic Review