A comparative analysis of 21 literature search engines

Acharya KK*, Greta K and Haritha H
IBAB, G-05, TechPark Mall, ITPB, Bangalore – 560 066, Karnataka State, India
**Corresponding author: kshitish@ibab.ac.in*

ABSTRACT:

With increasing number of bibliographic software, scientists and health professionals either make a subjective choice of tool(s) that could suit their needs or face a challenge of analyzing multiple features of a plethora of search programs. There is an urgent need for a thorough comparative analysis of the available bio-literature scanning tools, *from the user's perspective*. We report results of the first time semi-quantitative comparison of 21 programs, which can search published (partial or full text) documents in life science areas. The observations can assist life science researchers and medical professionals to make an informed selection among the programs, depending on their search objectives.

Some of the important findings are:

1. Most of the hits obtained from Scopus, ReleMed, EBImed, CiteXplore, and HighWire Press were usually relevant (i.e., these tools show a better precision than other tools).
2. But a very high number of relevant citations were retrieved by HighWire Press, Google Scholar, CiteXplore and Pubmed Central (they had better recall).
3. HWP and CiteXplore seemed to have a good balance of precision and recall efficiencies.
4. PubMed Central, PubMed and Scopus provided the most useful query systems.
5. GoPubMed, BioAsk, EBIMed, ClusterMed could be more useful among the tools that can automatically process the retrieved citations for further scanning of bio-entities such as proteins, diseases, tissues, molecular interactions etc).

The authors suggest the use of PubMed, Scopus, Google Scholar and HighWire Press - for better coverage, and GoPubMed - to view the hits categorized based on the MeSH and gene ontology terms.

INTRODUCTION:

Efficient search of published scientific articles is not only a key facilitator of the current speed of discoveries in life sciences, but also important for successful health management. Several databases and search engines *(see table 1)* have been created to enhance the efficiency of scanning published articles and retrieving the relevant citations[1-5]. But users face a new challenge with the increase in the number of novel tools: they now have to acquaint with multiple features of a plethora of search tools.

In this context, a systematic comparative study of different utilities of the available search tools would be helpful. Some studies have compared the search tools from a user's perspective. But such studies have considered very few search tools and often in the context of one specific domain[6-22]. A thorough application-based assessment of all major literature mining softwares, preferably a quantitative one, would help many scientists and physicians.

However, such a comparison is almost impossible. One of the main reasons for this difficulty is the diversity across the search engines. The existing literature search tools

can be categorized into 3 main types: (a) the simple summary-scanners, which are capable of searching for the key words only in the citations (title of the article and, author and journal details, with or without abstracts); (b) the full-text scanners, which can actually search the entire main-text of articles for the query terms/phrases; and (c) summary scanners and information processors, which can automatically process the retrieved citations to organize them in an useful way and/or extract further information. The tools also vary in the quality of the resources (of published literature) used, query flexibility allowed, search algorithms employed, presence and complexity of down-stream processing, and display of the output. There are other complications faced when one tries to compare these programs, including the possible variety of the search objectives.

Nevertheless, a semi-quantitative method can be used to evaluate the capacities and utilities across these search engines. We have taken such an approach to compare most of the commonly used tools, and rated the relative recall and precision efficiencies, the quality of the query system, the output and other features of these programs.

METHOD for Semi-quantitative Comparison of Search Tools:

The features of literature search were compared under the following major categories:

a) Citation retrieval efficiencies: the ability to scan and retrieve relevant citations: Three simple sets of query terms were used uniformly, irrespective of the features of the input pages of the tools considered. Since it is difficult to directly assess the search efficiencies, a 'relative recall efficiency' and an 'indicative precision value' were calculated. The relevance of articles was assessed by reading a specific number of sample abstracts from the results of each search for each tool (supplementary notes 1 has the scoring system: http://resource.ibab.ac.in/LITsearch *- Note: Do NOT use www in the URL*).

Three topics chosen for this component of the study were: RNA binding proteins in the context of transcription initiation, alternative promoters in mice, and cell death in the context of liver toxicity (see supplementary notes 1 for query terms: http://resource.ibab.ac.in/LITsearch).

Based on the comparative assessment in all 5 categories, some of the most useful and unique tools were again tested for their citation retrieval efficiency with 3 specific biological objectives (related to microRNA and cancer; piRNA in non-testicular tissues; and quadruplex DNA structure and HIV; see supplementary notes 3 for details: http://resource.ibab.ac.in/LITsearch). In this round, the best possible query set was derived using the query features of each tool.

b) *Query system quality: the efficiency with which query terms can be used and/or combined.*
c) *Resource coverage: the number and types of scientific documents scanned by the search tool.*
d) *Output quality: the display features.*
e) *Miscellaneous: other features, including the duration for which the results can be stored.*

Specific parameters were identified in each of the last 4 categories (b to e) for a semi-quantitative comparison of the search engines. Preliminary studies determined a

'relative potential impact/importance' of every parameter on the quality of user's search process and the output. Based on this assessment, a 'maximum possible score' for the parameters was then decided for each parameter.

For example, a maximum score of 10 was assigned to the 'number of query terms or characters allowed' while the feature allowing 'phrase searching' had an upper limit of 3 and 'truncation' of key word feature was given the higher limit of 2**.** Similarly, the history option in the PubMed, which could significantly affect the overall search efficiency, was given a higher upper limit (12 points) than the feature of enabling the search without the Boolean operators as in askMEDLINE (1 point).

The actual score was then assigned based on the specific aspects of the paramter across the tools. To cite a case, while CiteXplore received 4 points (of the maximum 10) for allowing up to 500 characters in the query (as determined by different trials), ClusterMed scored 9.5 as it allowed up to 3000 characters. Within each of the major categories, the tools were finally ranked on the basis of the sum of scores for all parameters. The scoring system for query quality, coverage, output quality and the miscellaneous features are explained in detail in the supplementary tables 1, 2, 3 and 4, respectively (http://resource.ibab.ac.in/LITsearch).

RESULTS:

Relative recall and precision efficiencies: Scopus, ReleMed, EBImed, CiteXplore, and HWP revealed reliable precision in the output (see fig. 1a and supplementary table 5; http://resource.ibab.ac.in/LITsearch).

The full-text search engines dominated the top positions when comparing relative recall efficiency, with HWP topping the list. However, CiteXplore, which is not a full-text searcher, attained a distant second position (see fig. 1b and supplementary table 6 http://resource.ibab.ac.in/LITsearch). Relemed failed to retrieve many relevant citations from the resources.

HWP and CiteXplore showed a good balance of precision and relative recall efficiencies.

Query quality: Keeping the query-sets uniform was essential to compare the inherent retrieval capacities of the programs. But the actual relevance of the results can also be influenced remarkably by the extent to which a search engine would allow the user to set intelligent query terms, phrases and/or their combinations. In fact, the efficiency of the query input interface can be the most important part of a search engine.

PubMed and PMC scored well in all parameters related to query set designing including the flexibility of the search terms and phrases, available field selections, and query refinement (see fig. 2a). It should be noted despite the better quality of query system, PMC cannot compete with HWP or GS in terms of the final output as the latter have better coverage, recall and precision features.

Output quality: BioAsk and GoPubMed scored very high in the overall output quality (see fig. 2b). This was mainly because of their visualization features, the ability to display the bio-entities contained within the title/abstract of articles, statistical analysis and the ability to sort citations using multiple criteria. While the

visualization feature works for the top 500 hits in BioAsk, GoPubMed was able to efficiently sort and group the top 10,000 hits. BioAsk and GoPubMed, however, have other features that are mutually exclusive.

HWP, PubMed, Scopus and EBIMed scored high in the 'primary output features', which included: a) the total number of citations that are actually displayed, b) display of sentences or parts of sentences with query term(s), c) ability to display all abstracts at a time d) free full text status display, d) links to related articles and e) citation analysis of every hit.

The extent of coverage of scientific journals was not very different across the search engines except PMC, which covers only a small number of journals and GS, which frequently extracts citations from several non-PubMed journals (see supplementary notes 2 and 3; http://resource.ibab.ac.in/LITsearch).

Scopus, PubMed, PMC and BioAsk provide the best options to store the results among all.

The scores corresponding to each specific feature are listed in the supplementary tables 7a (query quality), 8a (output quality), 9 (coverage) and 10 (miscellaneous). Further details of the scoring for the features are available in the supplementary tables 7b to e (query quality) and 8b to h (output quality) (http://resource.ibab.ac.in/LITsearch).

A few other observations made during our studies are listed in Table 1.

CONCLUSION:

In addition to aiding the users in making better judgements while using the search tools, the review would aid search engine designers via identification of pitfalls in the currently available search engines. Such periodic assessments are essential in view of the growing number of literature search engines, particularly in absence of an 'ideal search engine'. The first time tangible account of the relative strengths and weaknesses of most of the available search engines reveals that no single search engine can be relied upon for a thorough search of all relevant citations and/or automatic retrieval of information from abstracts. With every tool capable of different coverage and/or offering unique feature(s), the process of selecting one or few appropriate tool(s) becomes difficult.

However, two suggestions can be made from the current study for general biological literature searches:

a) For reasonable net retrieval efficiency, it may be better to derive a comprehensive non-redundant list of relevant citations from the results of the following tools: PubMed, Scopus, GS and HWP (see supplementary notes 3 for examples and a comparison of all the 3 free tools; http://resource.ibab.ac.in/LITsearch*).*

While HWP and GS can scan full documents for the query terms and cover different resources of documents, they often need repeated searches due to limitations in the query options (see supplementary notes 3 for examples). With GS, one may have to sometimes deal with high number of unwanted hits. Despite these limitations, GS can be used to gather more number of relevant citations, but with lots of extra work. In some searches, 60 to 70% of the relevant citations were contributed exclusively by the GS.

b) If one wants to analyze a large number of citations in the context of the bioentities contained in them, it helps to first use PubMed to arrive at the best combination of key words/phrases, and then apply the query set to GoPubMed (see supplementary notes 3). The number of abstracts that can be processed is high in GoPubMed and EBIMed (10,000) and ClusterMed (5000), unlike several other citation retrievers capable of automatic bioentity-based clustering of abstracts or further processing. The method of categorizing the citations is different in these tools.

References:

Note:

a) More references/suggested readings in supplementary notes 2

b) While connecting to the supplementary materials online, please d*o NOT use www in the URL. Copy & paste the following in the address bar of your browser:* http://resource.ibab.ac.in/LITsearch

1. Grivell, L., Mining the bibliome: searching for a needle in a haystack? New computing tools are needed to effectively scan the growing amount of scientific literature for useful information. *EMBO rep.* **3**, 200-203 (2002).
2. Al-Ubaydli M. Using search engines to find online medical information. *PLoS Med*. **2**(9):e228. Epub (2005);
3. Westbrook, J.I., Coiera, E.W. & Gosling, A.S., Do online information retrieval systems help experienced clinicians answer clinical questions? *J. Am. Med. Inform. Assoc*. **12**, 315–321 (2005).
4. Rodgers RP. Searching for biomedical information on the World Wide Web. *J Med Pract Manage.* **15**, 306-13 (2000).

   Soualmia LF, Dahamna B, Thirion B, Darmoni SJ. Strategies for health information retrieval. *Stud Health Technol Inform* 124:595-600 (2006).
5. Jensen, L.J., Saric, J. & Bork, P., Literature mining for the biologist: from information retrieval to biological discovery. *Nat. Rev. Genet.* **7,** 119-129 (2006).
6. Burrows, S.C., & Tylman, V., Evaluating medical student searches of MEDLINE for evidence-based information: process and application of results. *Bull. Med. Libr. Assoc.* **87**, 471–476 (1999).
7. Zhou, W., Smalheiser, N.R., & Yu, C., A tutorial on information retrieval: basic terms and concepts. *J. Biomed. Discov. Collab.* **1**, 2 (2006).
8. Ilic, D., Bessell, T.L., Silagy, C.A.., & Green, S., Specialized medical search-engines are no better than general search-engines in sourcing consumer information about androgen deficiency. *Hum. Reprod.* ***18,*** 557-561 (2003).
9. Bakkalbasi, N., Bauer, K., Glover, J., & Wang, L., Three options for citation tracking: Google Scholar, Scopus and Web of Science. *Biomed. Digit. Libr.* **3**, 7 (2006);

   Vanhecke TE, Barnes MA, Zimmerman J, Shoichet S. PubMed vs. HighWire Press: a head-to-head comparison of two medical literature search engines. *Comput Biol Med.* **37**, 1252-8 (2007).

   Yu H, Kaufman D. A cognitive evaluation of four online search engines for answering definitional questions posed by physicians. Pac Symp Biocomput. 328-39 (2007).

10. Pospisil, P., Iyer, L.K., Adelstein, S.J., & Kassis, A.I., A combined approach to data mining of textual and structured data to identify cancer-related targets. *BMC Bioinformatics* **7**, 354 (2006).
11. Vincent, B., Vincent, M. & Ferreira , C.G., Making PubMed searching simple: learning to retrieve medical literature through interactive problem solving. *Oncologist* **11**, 243-251 (2006).
12. Ripple, A.S., Expert googling: best practices and advanced strategies for using google in health sciences libraries. *Med. Ref. Serv. Q.* **25**, 97-107 (2006).
13. McGowan, J. & Sampson, M., Systematic reviews need systematic searchers. *J Med, Libr. Assoc.* **93**, 74–80 (2005).
14. Gruppen, L.D., Rana, G.K. & Arndt, T.S., A controlled comparison study of the efficacy of training medical students in evidence-based medicine literature searching skills. *Acad. Med.* **80**, 940-944 (2005).
15. Saxton, J.D. & Owen, D.J., Developing optimal search strategies for finding information on herbs and other medicinal plants in MEDLINE. *J. Altern. Complement Med.* **11,** 725-731 (2005).
16. Mijnhout, G.S., *et al*, How to perform a comprehensive search for FDG-PET literature. *Eur. J. Nucl. Med.* **27**, 91-97 (2000).
17. Robinson, K.A. & Dickersin, K., Development of a highly sensitive search strategy for the retrieval of reports of controlled trials using PubMed. *Int. J. Epidemiol.* **31***,* 150-153 (2002).
18. Falagas, M.E., Pitsouni, E.I., Malietzis, G.A. & Pappas G. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. FASEB J. **20,** 338-42 (2008).
19. Fremer, E. & Larsson, B., SPIRS, WinSPIRS, and OVID: a question of free-text versus thesaurus retrieval? Bull Med Libr Assoc. **85**, 57-8 (1997).
20. Schoonbaert, D., SPIRS, WinSPIRS, and OVID: a comparison of three MEDLINE-on-CD-ROM interfaces. Bull Med Libr Assoc. **84**,63-70 (1996).
21. Jacso, P., As we may search: comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. *Current Science*. **89**, 1537–47 (2005).
22. Shultz, M., Comparing test searches in PubMed and Google Scholar *J Med Libr Assoc.* **95**, 442–445 (2007).
23. Fontelo, P., Liu, F. & Ackerman, M., askMEDLINE: a free-text, natural language query tool for MEDLINE/PubMed. *BMC Med. Inform. Decis. Mak.* **5**, 5 (2005).
24. Lewis, J., Ossowski, S., Hicks, J., Errami, M. & Garner, H.R., Text similarity: an alternative way to search MEDLINE. *Bioinformatics* **22**, 2298-2304 (2006).
25. Errami, M., Wren, J.D., Hicks, J.M. & Garner, H.R., eTBLAST: a web server to identify expert reviewers, appropriate journals and similar publications. *Nucleic Acids Res.* **35**, W12-W15 (2007).
26. Muin, M., Fontelo, P., Liu, F. & Ackerman, M., SLIM: an alternative Web interface for MEDLINE/PubMed searches - a preliminary study. *BMC Med. Inform. Decis. Mak.* **5**, 37 (2005).
27. Muin, M. & Fontelo, P., Technical development of PubMed interact: an improved interface for MEDLINE/PubMed searches. *BMC Med. Inform. Decis. Mak.* **6**, 36 (2006).
28. Siadaty, M.S., Shu, J. & Knaus, W.A., ReleMed: sentence-level search engine with relevance score for the MEDLINE database of biomedical articles. *BMC Med. Inform. Decis. Mak.* **7**, 1 (2007).
29. Plake, C., Schiemann, T., Pankalla, M., Hakenberg, J. & Leser, U., AliBaba: PubMed as a graph. *Bioinformatics*, **22,** 2444-2445 (2006).
30. Divoli, A. & Attwood, T.K., BioIE: extracting informative sentences from the biomedical literature. *Bioinformatics* **21**, 2138-2139 (2005).

31. Kim, J.J. & Park, J.C., BioIE: retargetable information extraction and ontological annotation of biological interactions from the literature. *J. Bioinform. Comput. Biol.* **2**, 551-568 (2004).
32. Rebholz-Schuhmann D. *et al*, EBIMed—text crunching to gather facts for proteins from MEDLINE. *Bioinformatics* **23**, e237-e244 (2007).
33. Rebholz-Schuhmann D, *et al*, Protein annotation by EBIMed. *Nat. Biotechnol.* **24**, 902-903 (2006).
34. Doms, A. & Schroeder, M., GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res*., **33**, W783-W786, (2005).
35. Plikus, M.V., Zhang, Z. & Chuong, C.M., PubFocus: semantic MEDLINE/PubMed citations analytics through integration of controlled biomedical dictionaries and ranking algorithm. *BMC Bioinformatics* **7**, 424 (2006).
36. Ding, J., Hughes, L.M., Berleant, D., Fulmer, A.W., Wurtele, E. S., PubMed Assistant: a biologist-friendly interface for enhanced PubMed search. *Bioinformatics* **22**, 378–380, (2006).
37. Perez-Iratxeta, C., Bork, P. & Andrade, M.A., XplorMed: a tool for exploring MEDLINE abstracts. *Trends Biochem. Sci.* **26**, 573-575 (2001).
38. Perez-Iratxeta, C., Bork P. & Andrade M.A., Exploring MEDLINE abstracts with XplorMed. *Drugs Today (Barc)*. **38**, 381-389 (2002).
39. Perez-Iratxeta, C., Pérez AJ, Bork P, Andrade MA. Update on XplorMed: A web server for exploring scientific literature. *Nucleic Acids Research*. **31,** 3866-3868 (2003).

Please refer to Supplementary notes 2 for more references and suggested related readings

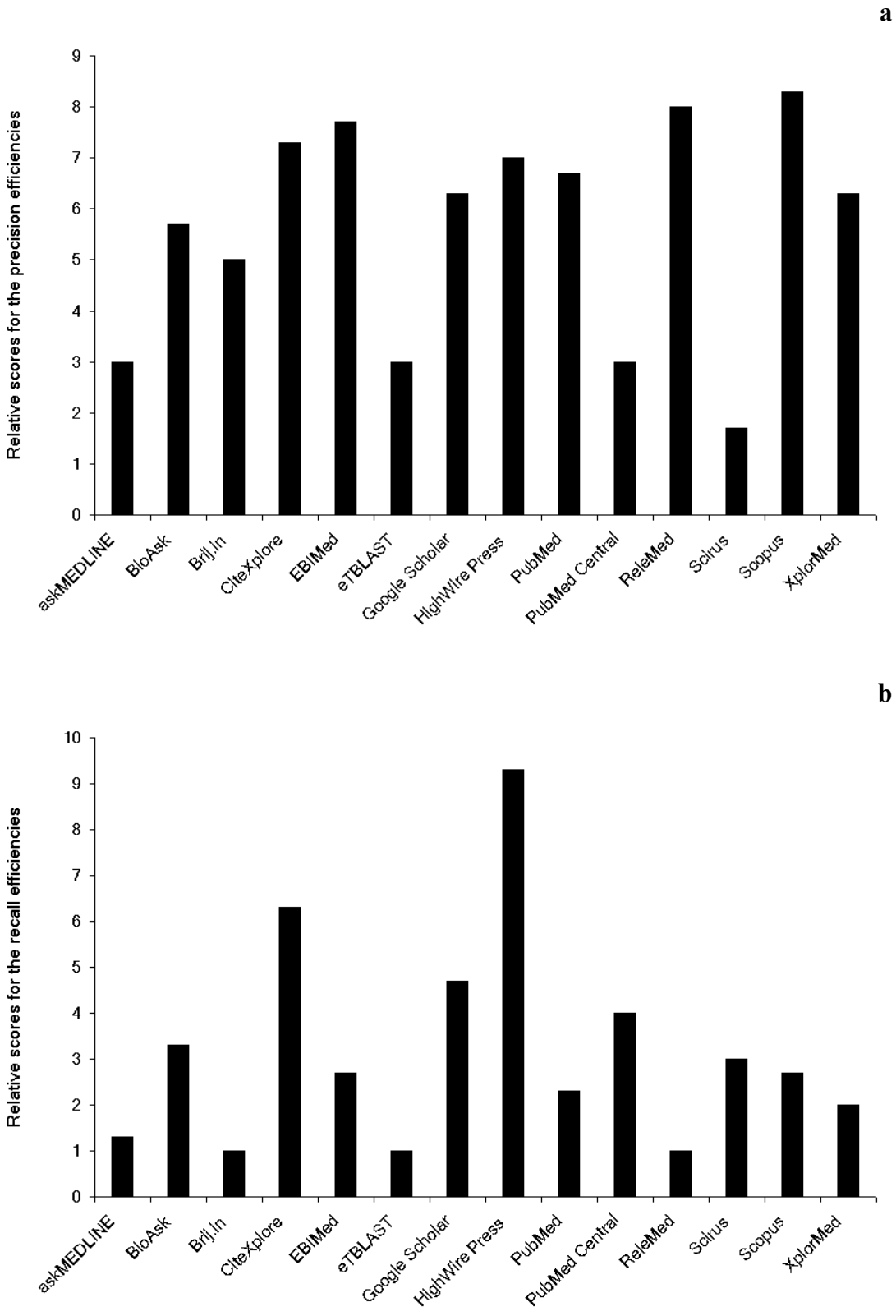

Figure 1. A comparison of indicative scores for indicative precision (**a**) and relative recall (**b**) of different search engines. Most of the tools that directly interface with PubMed were expected to have similar scoring as that of PubMed and hence, were not separately analyzed.

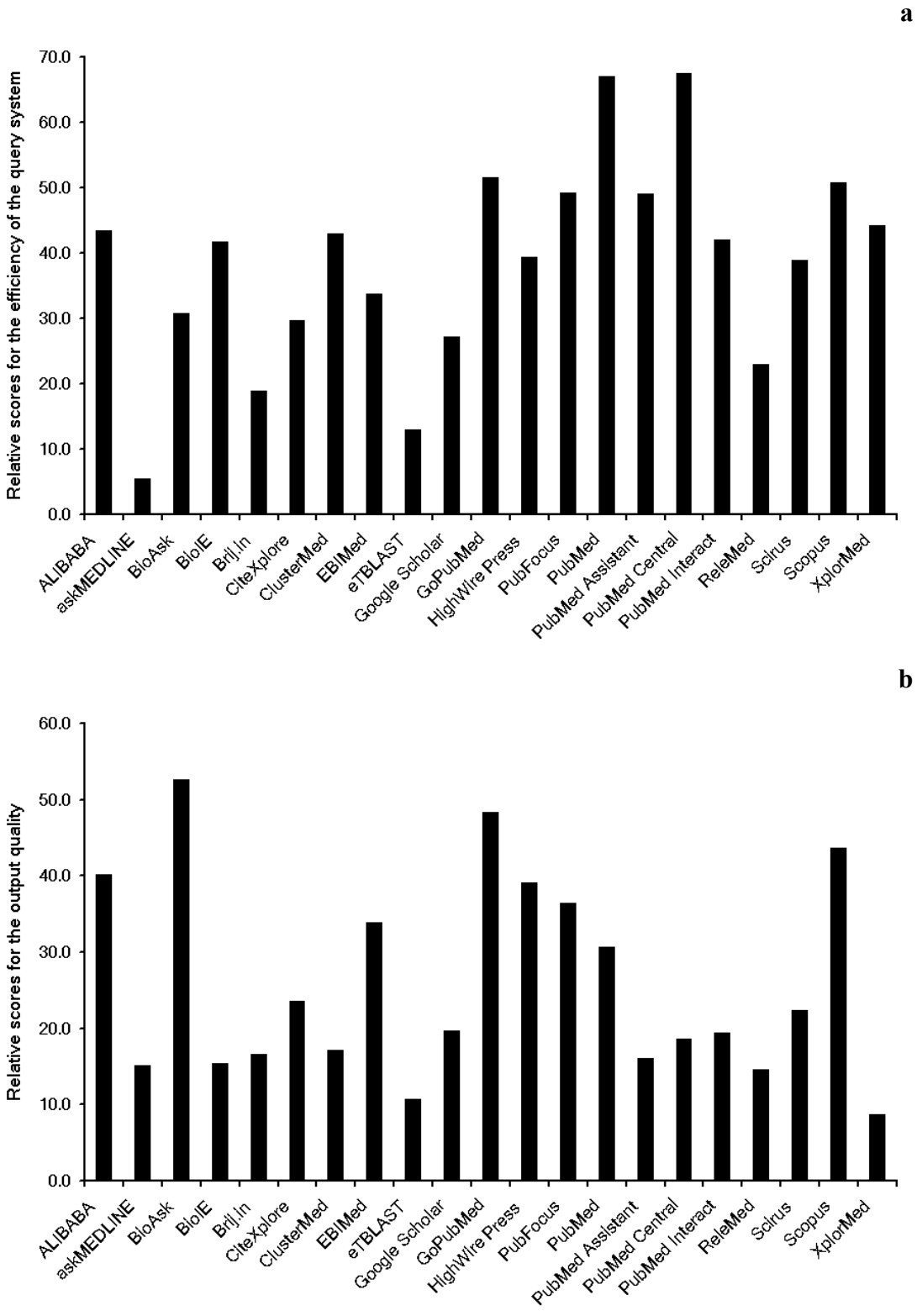

Figure 2. Results of the comparisons for two parameters, quality of the query system (a) and that of the output (b).

Table 1: The URLs and certain important observations on the literature search tools considered in the current study*.

Search tools	Observations and comments
A. Simple summary scanners	
askMEDLINE** [23] http://askMEDLINE.nlm.nih.gov/ask/ask.php	Easy but inefficient query system.
eTBLAST** [24,25] http://invention.swmed.edu/etblast/index.shtml	Allows queries with abstracts rather than key words. Poor precision obtained with abstracts could be improved by editing the query abstract, by enriching the relevant query terms in the paragraph.
PubMed http://www.ncbi.nlm.nih.gov/sites/entrez/	Popular, time-tested and commonly used tool. A relatively lower recall efficiency but one of the best query systems combined with good precision efficiency. Also offers good storage of results. The 'limits' features, subsets cannot always be taken for granted!
PubMed Interact** [26,27] https://pmi.nlm.nih.gov/interact/	Setting the PubMed limits using the slider bars of this tool saved 30 to 50% of time when comparing multiple combinations.
ReleMed** [28] http://www.relemed.com/	Specially built for sorting articles based on relevance. Good precision but poor recall.
B. Simple full-text scanners	
Google Scholar (GS) http://scholar.google.com/	Best for quick results, particularly for searching through the full-text articles. Low precision combined with high number of hits and limited query modulation features can form major set backs. Very good coverage. Displays less than 1000 results, irrespective of the total number of hits.
HighWire Press (HWP) http://highwire.stanford.edu/	Excellent recall and precision performances along with reasonably good quality query and output systems.
PubMed Central (PMC) http://www.PubMedcentral.nih.gov/	One of the best query systems. Capable of scanning full-text articles but from a limited set of resource journals.
Scirus http://www.scirus.com/	Impressive coverage of resources for searching but low precision in the output pages.
Life Science Search Engine http://www.brij.in	Though the tool focuses only on life science articles, we found the GS and many other tools better than this in many aspects.
C. Summary scanners that can process the citations for further information	
ALIBABA** [29] http://alibaba.informatik.hu-berlin.de/	One of the best visualization systems but very slow when processing more than 50 citations.
BioAsk http://www.bioask.com/	Has interesting novel features. Well-designed output system where more than 1300 hits are displayed. But the useful categorization is restricted to the top 500 hits.
BioIE** [30,31] http://www.bioinf.manchester.ac.uk/dbbrowser/bioie/	Slowed down when processing more than 400 articles.
CiteXplore http://www.ebi.ac.uk/citexplore/	Showed remarkable recall and precision but an average quality query system.
ClusterMed** http://demos.vivisimo.com/clustermed	Useful clustering of the results based on authors, affiliation, publication dates, MeSH terms etc.
EBIMed [32,33] http://www.ebi.ac.uk/Rebholz-srv/ebimed/index.jsp	Relatively slow in retrieving results. Also allows a very low number (20) of query terms.
GoPubMed** [34] http://www.goPubMed.org/	Excellent query and output systems.
PubFocus** [35] http://www.pubfocus.com/	Offers excellent statistical analysis of results and sorting of the results based on various parameters. Very slow when processing more than 50 results.
PubMed Assistant** [36] http://metnet.vrac.iastate.edu/browser/	Often slow in responding.
Scopus http://www.scopus.com/scopus/home.url	Perhaps the best coverage of resources. Also exhibited excellent precision in retrieving citations
XplorMed** [37-39] http://www.ogic.ca/project	Offers flexibility in setting the query; provides good recall but low precision.

*Since the studies involved extensive manual evaluations of citations and features, the number of tools selected was limited. Several tools not analyzed in detail in the current study for various reasons. For example, Hubmed, PubReminer, ConceptLink and PubMed Gold were very slow or didn't respond at all during our attempts to use them. Several other text mining tools are listed in the 'supplementary notes 2' http://resource.ibab.ac.in/LITsearch, which also provides other related information about tools listed above (sections B & C).

**Tools that interfaced with PubMed in an attempt to provide specific advantages.