Knowledge Discovery in Medline and Other Databases

Neil R. Smalheiser, MD, PhD

University of Illinois at Chicago
Smalheiser@psych.uic.edu

Overview

All neuroscientists are in the business of discovering knowledge about how the brain works. However, only a portion of time is spent in making new discoveries in the laboratory or clinic. An increasingly large task is to learn what has already been reported in the literature: either to assess an hypothesis and to plan out the best way to test it, or to keep abreast of new research trends, or simply to avoid rediscovering something already known. The days are gone when a person could keep up in neuroscience simply by scanning the pages of a few leading journals, or even by using alerting services such as Current Contents. Investigators not only need to become sophisticated users of Medline, the primary repository of published biomedical literature -- more than that, they need to go beyond simple queries. Think of getting information in Genbank: A simple query will retrieve the nucleotide sequence for "reelin", but one cannot directly look up the most probable transcription factor binding sites within its promoter region. Rather, specialized algorithms are needed to process the raw data and make plausible inferences (and these still need to be confirmed in the laboratory). Similarly, you can look up lots of findings in the biomedical literature, but to find knowledge that is implicit (not explicitly stated) and to make inferences, specialized approaches are needed. The purpose of this chapter is to guide neuroscientists in using informatics tools for making inferences in Medline as well as other public and private research databases.

I. What exactly is Text Data Mining? An example from Medline.

All of you probably know that Medline is a compendium of summaries of biomedical papers that have been published since 1966 in a core set of scientific journals screened for quality and relevance. However, you may not know that besides indexing fields such as the title, authors, journal, and abstract, each paper in Medline is read in its entirety by a professional biologist who assigns a set of terms called Medical Subject Headings (MeSH), which are chosen from a standard vocabulary. These terms describe what the paper is "really" about. Since the terms are not only standardized but also related to each other in a hierarchical fashion, one can search Medline for papers on a given topic by using MeSH. To search for a set of papers in PubMed, one searches among one or more Medline fields using a set of terms (and some options such as AND, OR, NOT and phrases " " or wildcard *). Let's say we want to find out information on dopamine D2 receptor expression in the adult rat brain. Typing in "dopamine D2 receptor" AND adult rat brain, PubMed gives a list of articles – not ranked in terms of importance, relevance or impact, by the way, and not clustered into sets of related articles, but simply listed in reverse chronological order.

Thus, Medline, and its query interfaces such as PubMed and Ovid, have been designed for people seeking to retrieve comprehensively all relevant papers on a given topic. [Exception: for clinical queries, there is also an option to retrieve the most relevant papers.] On the other hand, Medline does not bother to index other basic information related to authors: first initials instead of first names are given for authors (this is beginning to change in 2003), and affiliations are only recorded for the first author on a paper. The point here is to emphasize that query interfaces make it easy to search for **some** kinds of information, for **certain** purposes. But one cannot even pose certain basic questions regarding authors via the existing query interfaces: "Show me all of the papers on dopamine D2 receptors written by a sole author," or "all papers where Goldman-Rakic was listed as last author." Or "papers written by a particular individual, Rob W. Williams." There are many different people with the name RW Williams or even Robert W. Williams (and recall that middle initials are sometimes missing from papers, too). Knowing a person's affiliation is not sufficient to pin down an individual either – Rob Williams was first at Yale, then at U Tennessee, but he also is listed as co-author on papers whose affiliation fields are given as Oregon, Alabama and other institutions.

So, the task of finding papers written by a specific individual can be used here as a straightforward example of information that is not explicitly encoded within Medline, and that calls for some sophisticated large-scale data mining. Notice that the query interface is a hindrance rather than a help, because we need to take the relevant information out of the Medline records and put them into a relational database (briefly, a series of tables with rows and columns as entries), and the manner in which this is done depends on the algorithms you choose to employ. Vetle Torvik and I developed a statistical model in which two different papers (sharing the same author last name and first initial) are compared for similarity on 8 different aspects of the Medline record: the number of co-authors in common, the journal, the language used, the number of title words in common, the number of MeSH terms in common, number of affiliation words in common, and presence and match of middle initial and suffixes (e.g. Jr. or III). In order to do this, we had to encode these Medline fields in a manner that could readily be compared for a pair of papers. Thus, each pair of papers has a corresponding 8-dimensional comparison vector. Then, we developed two large reference sets: the **match set** consisted of many thousands of pairs of papers written by the same person, and the **non-match set** comprised many pairs of papers written by different individuals. For each reference set, we plotted the distribution of 8-dimensional comparison vectors that were observed. Then, to assess any query pair of papers that a person might want to compare, we calculate its 8-dimensional comparison vector, and see how often that vector occurs in the match set vs. in the non-match set. If this vector occurs much more frequently in the match set, the probability is high that both members of the query pair were written by the same individual. (Torvik et al., submitted for publication (1).) Finally, to permit people to submit queries, we have built a specialized query interface (the Author-ity tool, http://arrowsmith.psych.uic.edu), thus closing the circle.

II. Beyond Simple Queries: Assessing Hypotheses and Making Inferences

The above example was certainly mining **data**. Can one use text data mining to discover significant **knowledge**? Despite a lot of ongoing effort, computer algorithms have not yet been developed that can do more than recognize the simplest relationships and make the simplest inferences, based on assertions made in the text of scientific papers. For example, given the statement "NMDA receptor activation induces fos activity in the amygdala" a computer might be able to make the inference that "N-methyl D-asparate stimulates fos," and possibly make the generalization that "glutamate stimulates fos." On the other hand, the scientific mind regularly makes leaps and jumps that would make a salmon proud. (A falling apple leads to the idea of gravity.) Scientists readily make connections across disparate disciplines or arenas but currently this is done haphazardly. The idea is that computer-based tools being developed in the Arrowsmith project should enable scientists to find new knowledge, i.e., make discoveries, more rapidly, systematically, and comprehensively, than they could do on their own (2, 3). The discovery of new knowledge can refer to: a) discovering information already in the literature (that the scientist was simply unaware of); b) information that is not explicitly stated in the literature, but for which different separate pieces of evidence can be put together to support a plausible new inference; and c) new discoveries made in the laboratory or clinic. It is intended that the Arrowsmith project will stimulate all three kinds of discoveries.

Although the Arrowsmith website is still under construction, it is free, public, and fully functional, and can be viewed as extending PubMed searching to another dimension (fig. 1). That is, the searcher formulates two different PubMed searches, defining two different literatures "A" and "C" that may not overlap but that are hypothesized to be related in some way. Then, the computer compiles a list of all words and phrases that are found in the titles of each set (throwing away common and uninformative words), and displays the terms "B" that are in both sets. Each B-term represents an item or concept that might possibly link the two literatures. By filtering the list of B-terms to a manageable number of prime candidates, one can view the AB titles juxtaposed to the BC titles and decide whether they appear to indicate a biologically relevant relationship or inference. If so, then further literature searching (and laboratory experiments!) may be warranted.

Here are two different examples of knowledge that can be discovered with the Arrowsmith approach:

First, consider a doctor who sees a patient with two distinctive clinical signs: retinal detachment and an aortic aneurysm. He wonders, what diseases are known which share both signs? A PubMed search on "retinal detachment AND aortic aneurysm" retrieves only a single article, dealing with a disease called fibromuscular dysplasia. But almost certainly there are other diseases that cannot be captured via a simple query of this sort. How about an Arrowsmith query? Literature A is "retinal detachment" (fig. 2), and literature C is "aortic aneurysm" (fig. 3). There are 741 terms on the "raw" B-list (fig. 4),

but we are only interested in names of diseases so we restrict the terms to the semantic category of "disorders/disease or syndrome" (fig. 5), leaving 103 terms that can be scanned quickly (fig. 6). Several connective tissue disorders are on the list (e.g., Marfan syndrome); so are several autoimmune diseases (e.g. lupus), and a number of different infections (e.g. tuberculosis). Most of the B-terms are actually valid examples of diseases known to be associated with both retinal detachment and aortic aneurysm. For example, fig. 7 shows that in the case of Ehler Danlos syndrome, the titles on the left directly imply that patients have an association with retinal detachment, whereas the titles on the right indicate that they also have an association with aortic aneurysm. (Ehlers-Danlos appears as "Ehler Danlo" because the software removes final "s" from B-terms, in order to merge singular and plural forms). By linking on the PubMed ID number above any title, one can view the abstract and investigate the paper in more detail. So why did a standard PubMed search not detect these examples? It is because few people write about both signs in the same paper; usually they write about one or the other in different contexts. Arrowsmith is at its best at putting together knowledge that is present in separate pieces and juxtaposing them so that they can be seen as fitting together.

Another example is the use of Arrowsmith to identify potentially "hot" research topics that are worth studying but that no one has yet published a paper on (e.g. 4-7). A few years ago, an epidemiologic paper reported an association between estrogen supplementation and protection from Alzheimer disease, suggesting that there is a mechanistic link between estrogen and AD. But what links are most likely to be relevant to AD, that have not already been studied (and published on)? We carried out a search with literature A = estrogen and literature C = Alzheimer disease. By examining the B-terms that represented physiologic effects with the corresponding AB and BC papers, we identified a short-list of 8 potential links. Most significantly of these, a substantial set of papers indicated that estrogen exhibits antioxidant activity, and a substantial literature reported that oxidative damage occurs in AD at the cellular level. Thus, we suggested that a promising avenue of research would be to test whether estrogen's antioxidant activity was relevant to its protective effect against AD. At the time this analysis was submitted for publication (5), no one had published such a test. However, several positive reports followed shortly thereafter, validating both the hypothesis and the fact that this was indeed a "hot" research topic (notwithstanding the fact that the association of estrogen and AD continues to be controversial and imperfectly understood).

The Arrowsmith tools are being developed with the support of a phase I Neuroinformatics grant from the Human Brain Project and the National Library of Medicine. Essential to this work are the efforts of several neuroscience research groups participating as field testers, who are assessing not only whether the tools work satisfactorily but whether they actually help bench scientists to formulate new lines of research, and (when indicated) to form new collaborations with others in different disciplines.

III. Beyond Simple Inferences: Linking Bio-Informatic and Clinical Databases

The concept of making AB-BC inferences across disparate literatures is not restricted to bibliographic databases such as Medline. Nor is one restricted to data that reside within a single database. If one database has data indicating a relationship between A and B, and another database indicates a relationship between B and C, then (depending on the particulars) one may be entitled to suggest that A is related to C -- even though A and C have not been measured together in the same study or in the same research subjects.

Perhaps the most promising example of how one might mine data across studies involves different inbred mouse lines and recombinant crosses thereof. These mouse lines have been characterized in terms of many behavioral phenotypes, their patterns of gene expression in microarrays, and various neuroanatomical parameters, and have been subjected to QTL analysis to map these phenotypes to chromosomal loci. If two different phenotypes (studied separately) vary together across strains, then one would like to be able to predict which of these are related mechanistically to each other. Going further, one would like to predict which genes or neural systems are most likely to underlie the correlations observed among phenotypes. The variation across studies is probably less than in any other field of neuroscience insofar as the mice are supposed to be genetically identical within each strain, so it may be acceptable to pool all of the available data and regard them as arising from one large study. Nevertheless, individual animals do differ in terms of age, gender, housing conditions, and environmental and dietary influences, so the results from different studies may not necessarily be comparable.

One might despair at trying to mine data across clinical studies at all, because of the great heterogeneity in human populations, differences in research protocols, and different methods for measuring the same basic parameter (for example, there are many different ways to measure “pain” or “obesity” that are not quite equivalent). However, it is impossible to collect from a single subject all of the data that is relevant to a disease such as schizophrenia, so each study can capture at best a single facet, a single piece of the puzzle. Data mining across studies is nothing more or less than the attempt to put the pieces together. This task can be helped by ensuring that all clinical studies include common “bridging” parameters that can be used to help calibrate studies against each other.

Certainly there are major challenges to data mining within a study (database) based on statistical correlations, and there are additional issues to confront if one wants to make inferences across different databases. For example, one must have access to the metadata to ensure that the studies in each database were conducted in a way that can be compared directly. The parameters A, B and C must be connected in some mechanistically meaningful way. And the nature of the relationship must be constrained such that the transitive inference makes sense (A-B and B-C must imply A-C). Finally, it is uncertain how to calculate the statistical significance of an A-C inference. However, as more and more scientists archive their primary research results in databases, and as data sharing becomes more and more common, then data mining across different databases will become an increasingly important endeavor (8, 9).

Conclusions

The major take-home lesson is that individual bench scientists can (and should) use a variety of informatics tools to assist their research. **First**, scientists can use text-based tools to gain more sophisticated access to published information in order to assess their hypotheses, and prioritize and design their experiments. Today, most investigators find this information haphazardly. **Second**, scientists need to envision and archive their experiments in a new way that can be summed up as: "One Study, One Database." I realize that there are practical difficulties in doing this at present, but this is the goal of much work in neuroinformatics, and I will be happy if all of you simply start picturing studies as databases. Putting primary experimental data in databases (along with the metadata that describe the experimental conditions) permits them to be analyzed not only via conventional hypothesis testing, but also by looking for statistical correlations within and across databases. Putting each experiment into database form may add limited value to the current way of doing science, but when thousands of experiments are pooled and pieced together, the overview can be remarkably coherent and reliable. (Think of how valuable expressed sequence tag (EST) databases have been in genomics, even though each individual EST by itself has very low quality.)

Finally, the informatics-savvy scientist recognizes that today's razor-sharp hypothesis is likely to be seen as ill-formed and even laughable 10 years from now, but data are forever. If one only collects and analyzes data that are strictly relevant to today's hypothesis (the "classic" view of experimental design), then one will lose the potential **future** value of the data to be reanalyzed in the light of other advances and other investigators in the field (9, 10).

Acknowledgments

This Human Brain Project/Neuroinformatics research is funded jointly by the National Library of Medicine and the National Institute of Mental Health.

Free, Public Links to Assist in Mining Text in the Biomedical Literature
(see the Arrowsmith site for a larger compendium of search tools)

PubMed
http://pubmed.gov
extremely popular site for searching Medline, that is linked to other NIH bio-informatic databases.

Arrowsmith
http://arrowsmith.psych.uic.edu
site for searching links between two literatures within Medline.
Also contains the Author-ity tool for disambiguating authors on scientific papers.

Pubcrawler
http://pubcrawler.ie
alerting service that searches custom queries in PubMed or Genbank automatically and notifies the user when new relevant papers or sequences appear in the literature.

Vivisimo
http://www.vivisimo.com
demo site that searches PubMed (or the Web) and arranges the output into clusters of articles that are most thematically similar to each other.

Medminer
http://discover.nci.nih.gov/textmining/filters.html
site for extracting information about gene-gene and gene-drug interactions from the abstracts of papers in Medline.

CompletePlanet
http://completeplanet.com
compendium of search engines, including U. S. Patents, Cancer Net, Census Bureau, Library of Congress, and many more.

Abbreviations server.
http://abbreviation.stanford.edu/
an online dictionary of abbreviations in PubMed articles.

References

1. Torvik, V. I., Weeber, M., Swanson, D. R. and Smalheiser, N. R. (2003) A probabilistic similiarity metric for Medline records: a model for author name disambiguation. J. Am. Soc. Information Sci. Technol, submitted.
2. Smalheiser, N.R. and Swanson, D.R. (1998) Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. Computer Methods and Programs in Biomedicine 57: 149-153.
3. Swanson, D.R. and Smalheiser, N.R. (1997) An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artificial Intelligence 91: 183-203.
4. Smalheiser, N.R. and Swanson, D.R. (1994) Assessing a gap in the biomedical literature: magnesium deficiency and neurologic disease. Neurosci. Res. Commun. 15: 1-9.
5. Smalheiser, N.R. and Swanson, D.R. (1996) Linking estrogen to Alzheimer's Disease: an informatics approach. Neurology 47: 809-810.
6. Smalheiser, N.R. and Swanson, D.R. (1996) Indomethacin and Alzheimer s Disease. Neurology 46: 583.
7. Smalheiser, N.R. and Swanson, D.R. (1998) Calcium-independent phospholipase A2 and schizophrenia. Arch. Gen. Psychiat. 55: 752-753.
8. Smalheiser, N.R. (2003) Linking investigators. A centralized linking facility for data sharing and coordination of samples in tissue banks. EMBO Reports 4: 108-110.
9. Smalheiser, N.R. (2002) Informatics and hypothesis-driven research. EMBO Reports 3: 702.
10. Koslow, S. H. (2000) Should the neuroscience community make a paradigm shift to sharing primary data? Nature Neurosci. 3: 863-865.

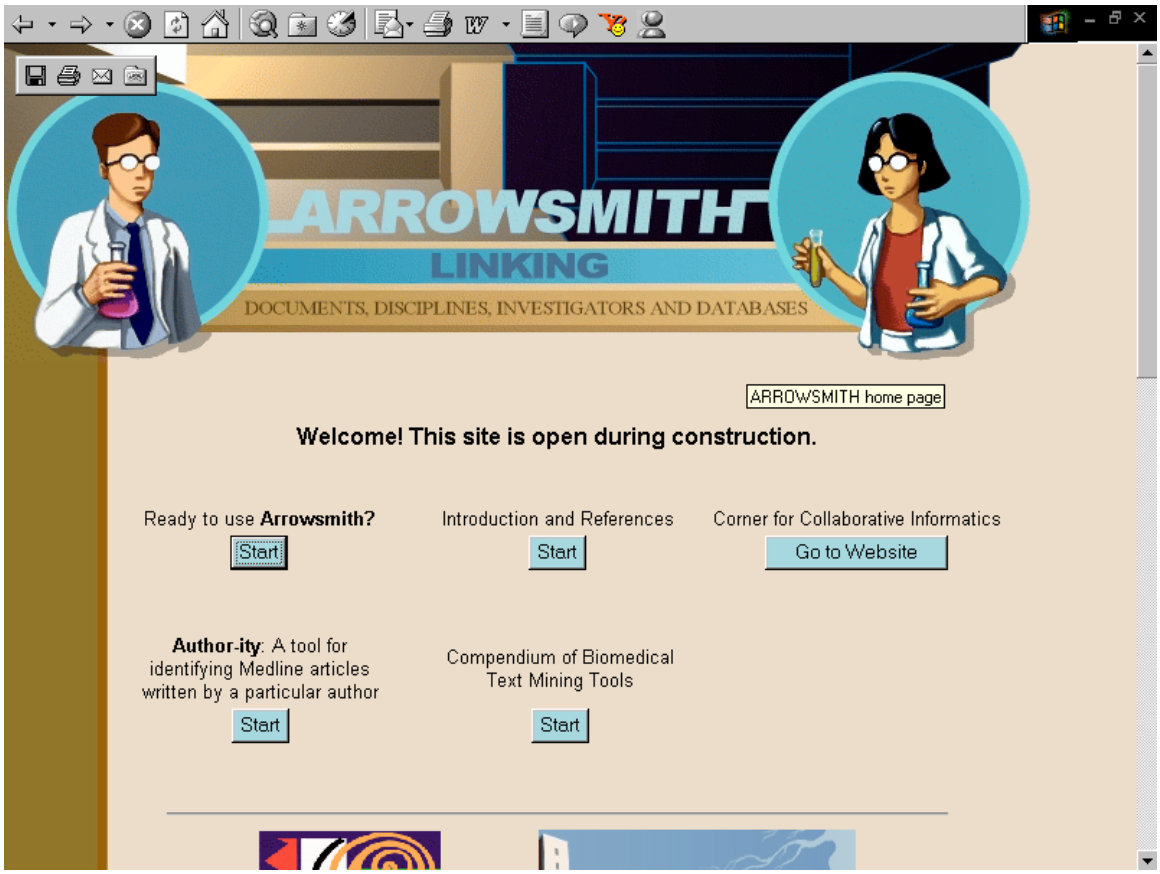
ARROWSMITH
LINKING
DOCUMENTS, DISCIPLINES, INVESTIGATORS AND DATABASES
ARROWSMITH home page
Welcome! This site is open during construction.
Ready to use Arrowsmith?
Start
Introduction and References
Start
Corner for Collaborative Informatics
Go to Website
Author-ity: A tool for identifying Medline articles written by a particular author
Start
Compendium of Biomedical Text Mining Tools
Start

Fig. 1

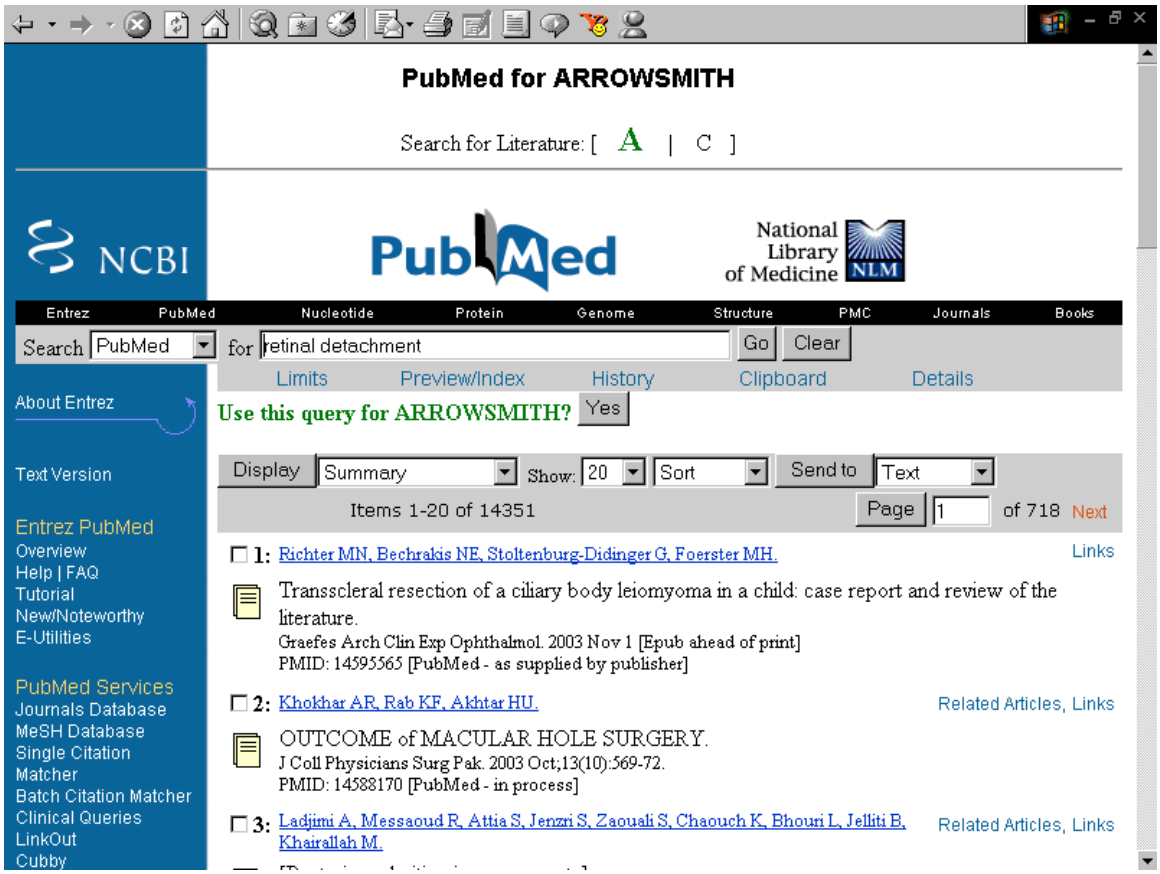
PubMed for ARROWSMITH
Search for Literature: [A | C]
NCBI
PubMed
National Library of Medicine
NLM
Entrez PubMed Nucleotide Protein Genome Structure PMC Journals Books
Search PubMed for retinal detachment Go Clear
Limits Preview/Index History Clipboard Details
About Entrez
Use this query for ARROWSMITH? Yes
Text Version
Display Summary Show: 20 Sort Send to Text
Items 1-20 of 14351 Page 1 of 718 Next
Entrez PubMed
Overview
Help | FAQ
Tutorial
New/Noteworthy
E-Utilities
PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
LinkOut
Cubby
1: Richter MN, Bechrakis NE, Stoltenburg-Didinger G, Foerster MH. Links
Transscleral resection of a ciliary body leiomyoma in a child: case report and review of the literature.
Graefes Arch Clin Exp Ophthalmol. 2003 Nov 1 [Epub ahead of print]
PMID: 14595565 [PubMed - as supplied by publisher]
2: Khokhar AR, Rab KF, Akhtar HU. Related Articles, Links
OUTCOME of MACULAR HOLE SURGERY.
J Coll Physicians Surg Pak. 2003 Oct;13(10):569-72.
PMID: 14588170 [PubMed - in process]
3: Ladjimi A, Messaoud R, Attia S, Jenzri S, Zaouali S, Chaouch K, Bhouri L, Jelliti B, Khairallah M. Related Articles, Links

Fig. 2

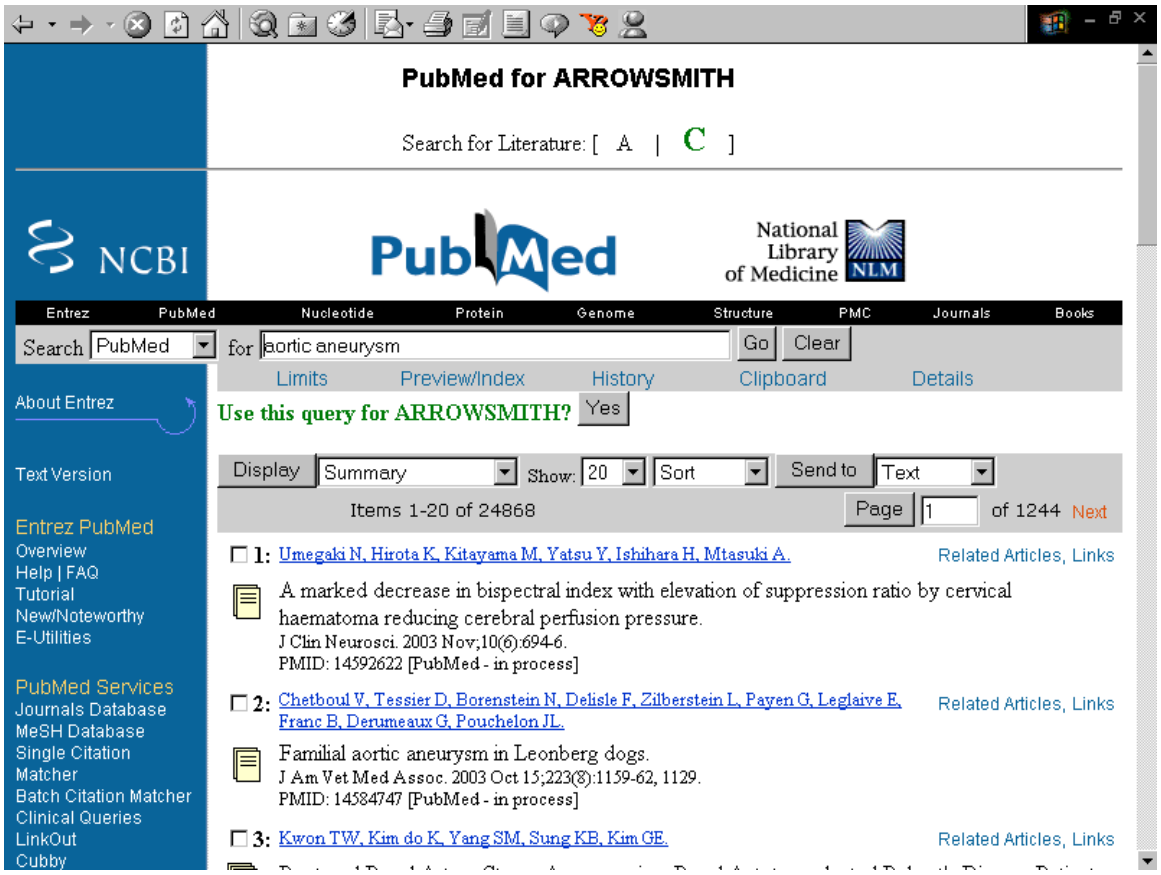
PubMed for ARROWSMITH
Search for Literature: [A | C]
NCBI
PubMed
National Library of Medicine
NLM
Entrez PubMed Nucleotide Protein Genome Structure PMC Journals Books
Search PubMed for aortic aneurysm Go Clear
Limits Preview/Index History Clipboard Details
About Entrez
Use this query for ARROWSMITH? Yes
Text Version
Display Summary Show: 20 Sort Send to Text
Items 1-20 of 24868
Page 1 of 1244 Next
Entrez PubMed
Overview
Help | FAQ
Tutorial
New/Noteworthy
E-Utilities
PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
LinkOut
Cubby
1: Umegaki N, Hirota K, Kitayama M, Yatsu Y, Ishihara H, Mtasuki A.
Related Articles, Links
A marked decrease in bispectral index with elevation of suppression ratio by cervical haematoma reducing cerebral perfusion pressure.
J Clin Neurosci. 2003 Nov;10(6):694-6.
PMID: 14592622 [PubMed - in process]
2: Chetboul V, Tessier D, Borenstein N, Delisle F, Zilberstein L, Payen G, Leglaive E, Franc B, Derumeaux G, Pouchelon JL.
Related Articles, Links
Familial aortic aneurysm in Leonberg dogs.
J Am Vet Med Assoc. 2003 Oct 15;223(8):1159-62, 1129.
PMID: 14584747 [PubMed - in process]
3: Kwon TW, Kim do K, Yang SM, Sung KB, Kim GE.
Related Articles, Links

Fig. 3

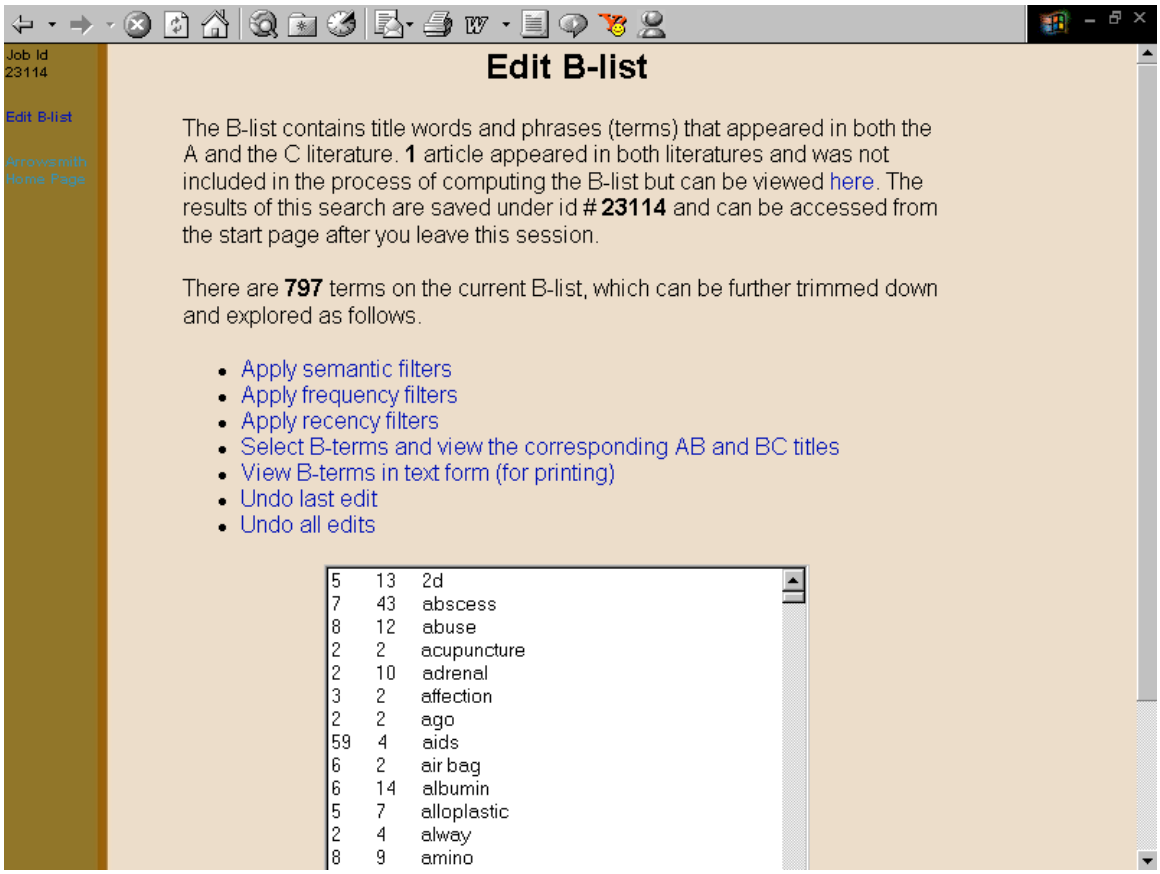
Job Id
23114
Edit B-list
Arrowsmith
Home Page
Edit B-list
The B-list contains title words and phrases (terms) that appeared in both the A and the C literature. 1 article appeared in both literatures and was not included in the process of computing the B-list but can be viewed here. The results of this search are saved under id # 23114 and can be accessed from the start page after you leave this session.
There are 797 terms on the current B-list, which can be further trimmed down and explored as follows.
Apply semantic filters
Apply frequency filters
Apply recency filters
Select B-terms and view the corresponding AB and BC titles
View B-terms in text form (for printing)
Undo last edit
Undo all edits
5 13 2d
7 43 abscess
8 12 abuse
2 2 acupuncture
2 10 adrenal
3 2 affection
2 2 ago
59 4 aids
6 2 air bag
6 14 albumin
5 7 alloplastic
2 4 alway
8 9 amino

Fig. 4

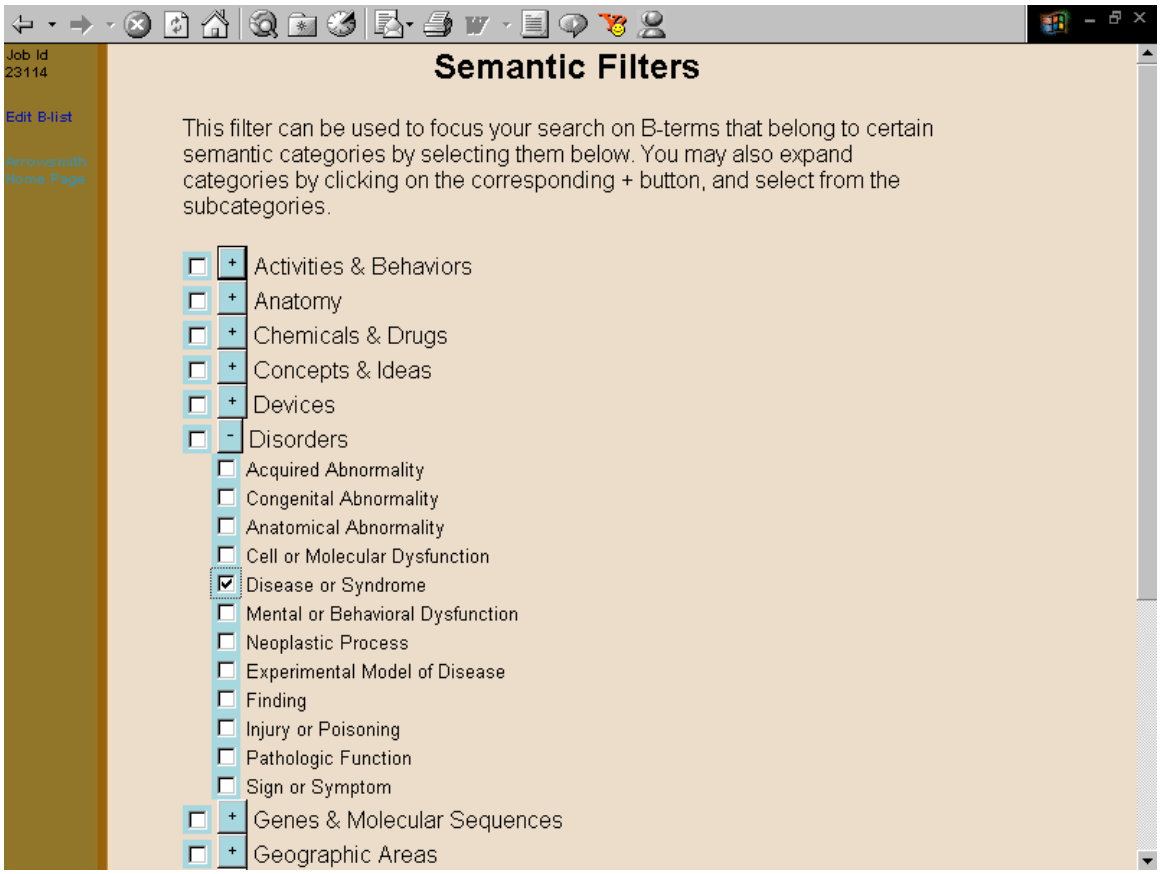
Job Id
23114
Edit B-list
Arrowsmith
Home Page
Semantic Filters
This filter can be used to focus your search on B-terms that belong to certain semantic categories by selecting them below. You may also expand categories by clicking on the corresponding + button, and select from the subcategories.
+ Activities & Behaviors
+ Anatomy
+ Chemicals & Drugs
+ Concepts & Ideas
+ Devices
- Disorders
Acquired Abnormality
Congenital Abnormality
Anatomical Abnormality
Cell or Molecular Dysfunction
Disease or Syndrome
Mental or Behavioral Dysfunction
Neoplastic Process
Experimental Model of Disease
Finding
Injury or Poisoning
Pathologic Function
Sign or Symptom
+ Genes & Molecular Sequences
+ Geographic Areas

Fig. 5

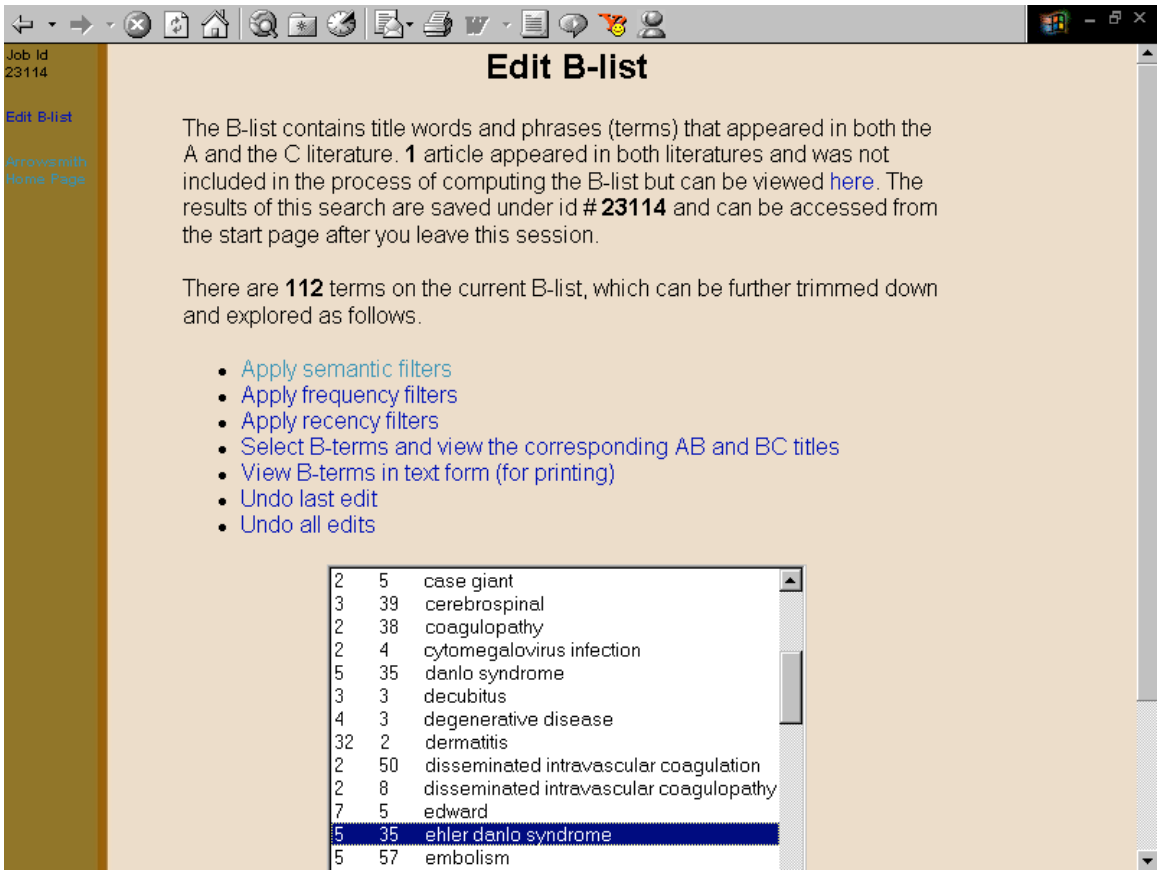
Job Id
23114
Edit B-list
Arrowsmith
Home Page
Edit B-list
The B-list contains title words and phrases (terms) that appeared in both the A and the C literature. 1 article appeared in both literatures and was not included in the process of computing the B-list but can be viewed here. The results of this search are saved under id # 23114 and can be accessed from the start page after you leave this session.
There are 112 terms on the current B-list, which can be further trimmed down and explored as follows.
Apply semantic filters
Apply frequency filters
Apply recency filters
Select B-terms and view the corresponding AB and BC titles
View B-terms in text form (for printing)
Undo last edit
Undo all edits
2 5 case giant
3 39 cerebrospinal
2 38 coagulopathy
2 4 cytomegalovirus infection
5 35 danlo syndrome
3 3 decubitus
4 3 degenerative disease
32 2 dermatitis
2 50 disseminated intravascular coagulation
2 8 disseminated intravascular coagulopathy
7 5 edward
5 35 ehler danlo syndrome
5 57 embolism

Fig. 6

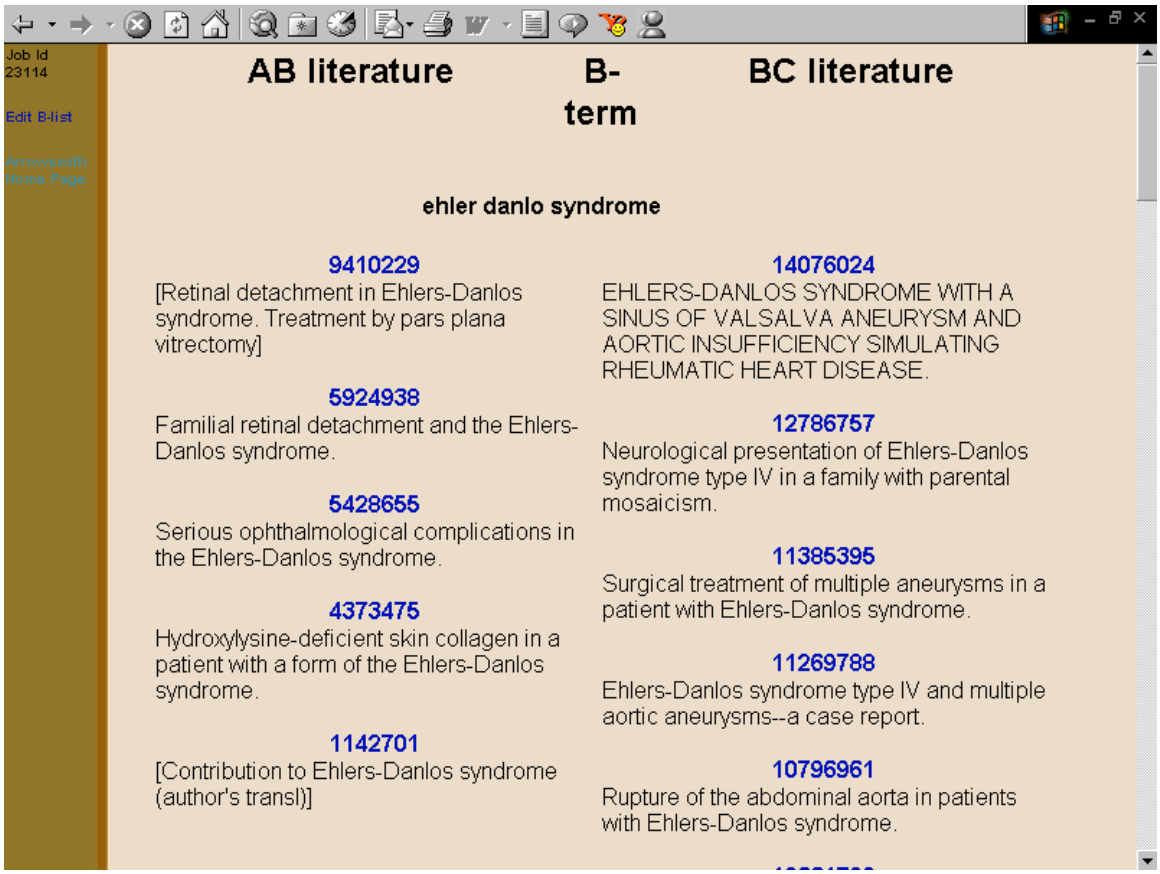
Job Id
23114
Edit B-list
Arrowsmith
Home Page
AB literature
B-term
BC literature
ehler danlo syndrome
9410229
[Retinal detachment in Ehlers-Danlos syndrome. Treatment by pars plana vitrectomy]
5924938
Familial retinal detachment and the Ehlers-Danlos syndrome.
5428655
Serious ophthalmological complications in the Ehlers-Danlos syndrome.
4373475
Hydroxylysine-deficient skin collagen in a patient with a form of the Ehlers-Danlos syndrome.
1142701
[Contribution to Ehlers-Danlos syndrome (author's transl)]
14076024
EHLERS-DANLOS SYNDROME WITH A SINUS OF VALSALVA ANEURYSM AND AORTIC INSUFFICIENCY SIMULATING RHEUMATIC HEART DISEASE.
12786757
Neurological presentation of Ehlers-Danlos syndrome type IV in a family with parental mosaicism.
11385395
Surgical treatment of multiple aneurysms in a patient with Ehlers-Danlos syndrome.
11269788
Ehlers-Danlos syndrome type IV and multiple aortic aneurysms--a case report.
10796961
Rupture of the abdominal aorta in patients with Ehlers-Danlos syndrome.

Fig. 7