Biomedical Information Retrieval

William Hersh, MD
Department of Medical Informatics & Clinical Epidemiology
School of Medicine
Oregon Health & Science University
hersh@ohsu.edu
www.billhersh.info

Searching – everyone is doing it …

"First, they do an on-line search."

OREGON HEALTH & SCIENCE UNIVERSITY

… everyone knows about it …

… but new problems have emerged

OREGON HEALTH & SCIENCE UNIVERSITY

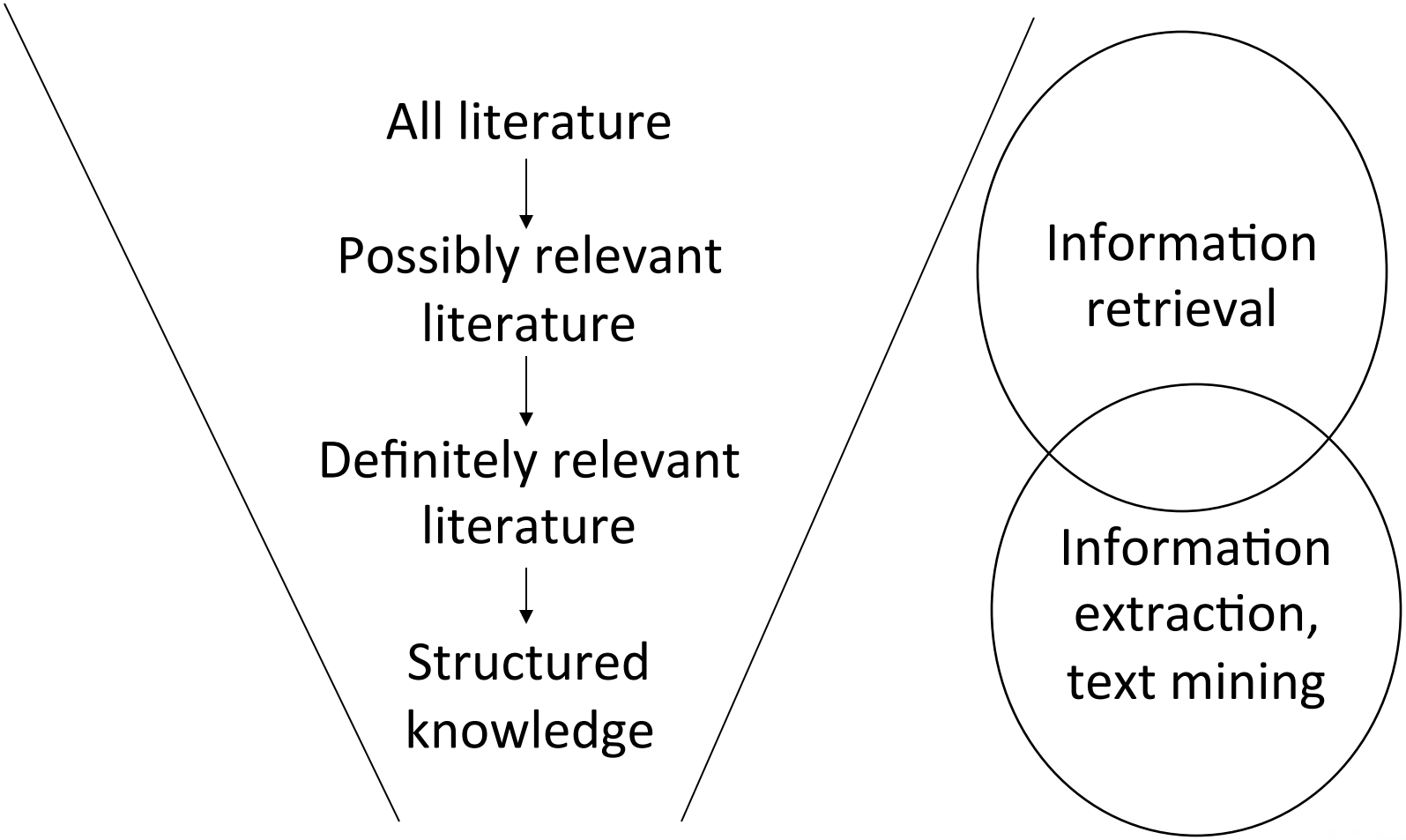
Biomedical information retrieval (IR)

1. IR in Biomedicine
2. Biomedical IR Content
3. Evaluation
4. Research Directions

IR in biomedicine: major challenges

- We have gone from
  - Information paucity to information overload
  - Paternalistic clinicians to engaged patients
  - Need to reduce waste in healthcare
- Many topics we want to search on have multiple ways to be expressed, e.g., diseases, genes, symptoms, etc.
- The converse is a problem too: Many words and terms used to express topics have multiple meanings
- Balancing open access vs. providing for cost of production and maintenance

IR is a growing part of "knowledge discovery" (Hersh, 2009)



Who uses biomedical IR systems?

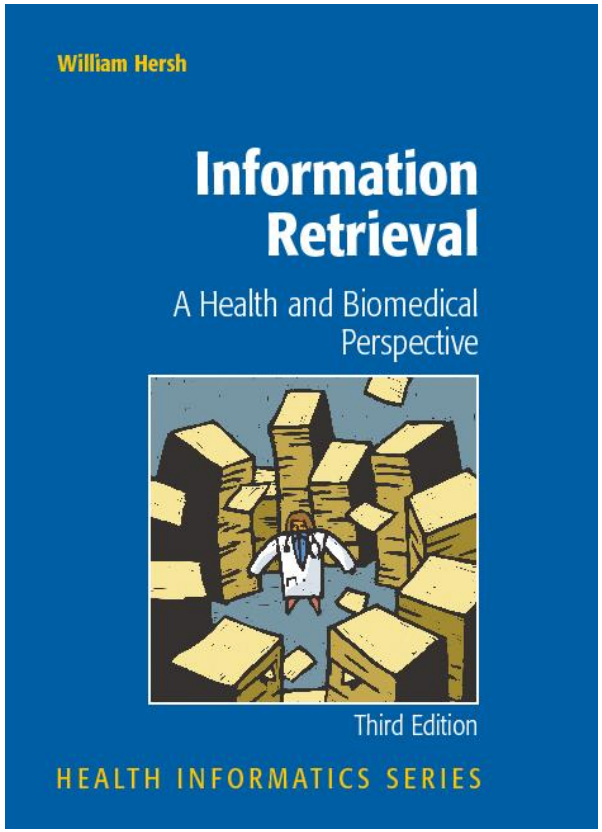
- Just about all Internet users "search" (if for no other reason than being sent to search pages when URLs fail)
- Most Internet users search for health information
  - Estimates for US adult Internet users who have searched for personal health information is about 80% (Taylor, 2011; Fox, 2011)
  - Virtually all US, Canadian, and UK physicians (and probably those from everywhere else) use electronic sources (Davies, 2010)
  - Large proportion of academic faculty (78-88%) and trainees (88%) own smartphones and use them for information access (Franko, 2011)

What kind of health information do people search for? (Fox, 2011)

Health topic	% searching
Specific disease or medical problem	66%
Certain medical treatment or procedure	56%
Doctors or other health professionals	44%
Hospitals or other medical facilities	36%
Health insurance – private or government	33%
Food safety or recalls	29%
Environmental health hazards	22%
Pregnancy and childbirth	19%
Medical test results	16%

How to find more information about biomedical IR

- From me!
- Hersh WR, *Information Retrieval: A Health and Biomedical Perspective*, Third Edition, 2009
  - Web site: www.irbook.info
- OHSU BMI 514/614 – Information Retrieval
- Many other good books, journals, and other sources as well

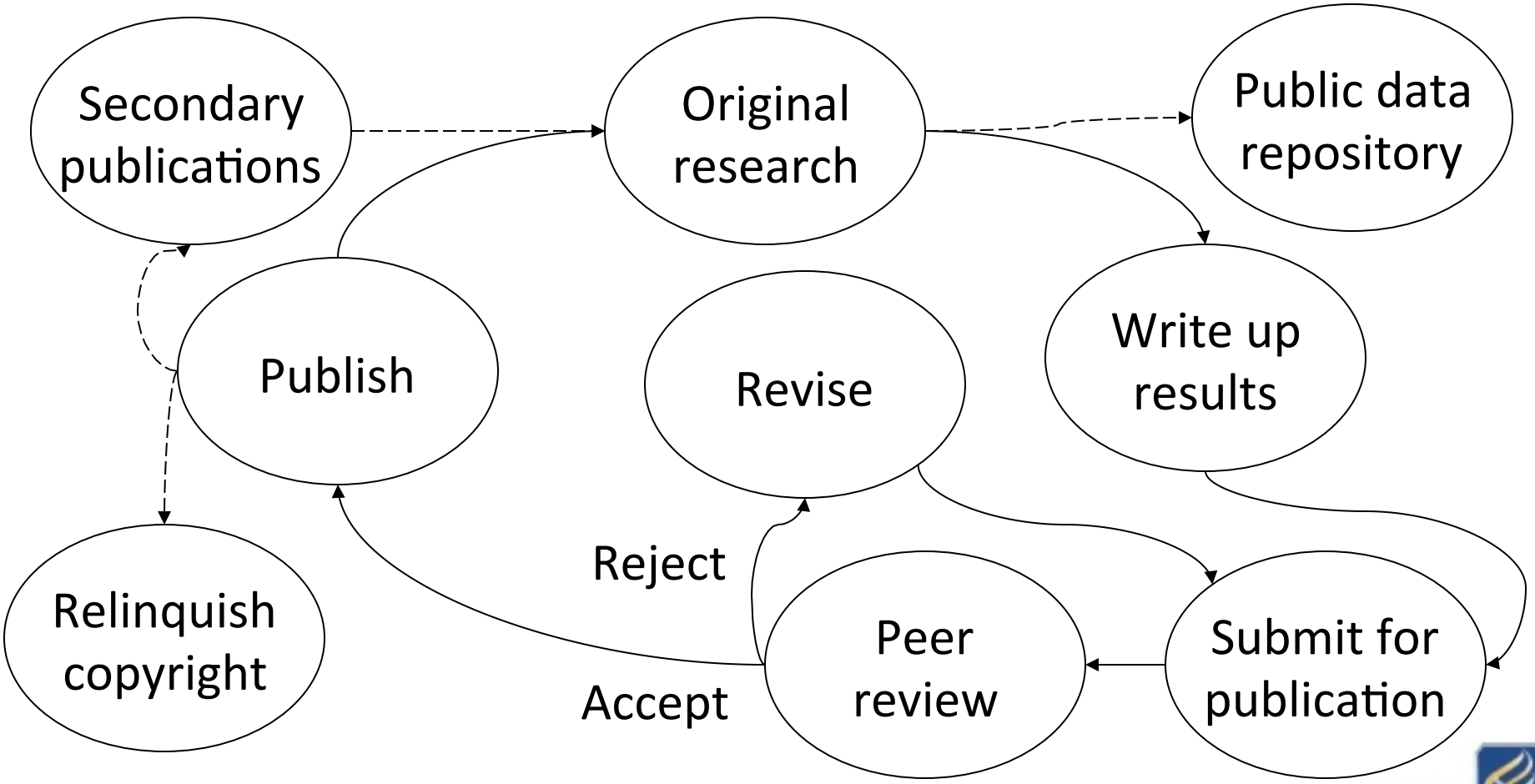


OREGON HEALTH & SCIENCE UNIVERSITY

Why is IR pertinent to health and biomedicine?

- Growth of knowledge has long surpassed human memory capabilities
- Clinicians have frequent and unmet information needs
- Researchers must frequently update their knowledge in new areas quickly
- Primary literature on a given topic can be scattered and hard to synthesize
- Non-primary literature sources are often neither comprehensive nor systematic
- Web is increasingly used as source of health and biomedical information

The life-cycle of knowledge-based information (Hersh, 2009)



Classification of knowledge-based scientific information

- Primary – original research
  - Published mainly in journals but also in conference proceedings, technical reports, books, etc.
  - Can include re-analysis, e.g., meta-analysis and systematic reviews
- Secondary – reviews, condensations, and/or synopses of primary literature
  - Textbooks and handbooks are staples of clinical practitioners, researchers, and others
  - Guidelines are important for normalizing care and measuring quality

Biomedical IR content: a classification

- Bibliographic
  - By definition rich in metadata
- Full-text
  - Everything on-line
- Annotated
  - Non-text or structured text annotated with text
- Aggregations
  - Bringing together all of the above
- These categories are admittedly fuzzy, and increasing numbers of resources have more than one type

Bibliographic content

- Bibliographic databases
  - The old (e.g., MEDLINE) have been revitalized with new features
  - New ones (e.g., National Guidelines Clearinghouse) have emerged
- Web catalogs
  - Share many characteristics of traditional bibliographic databases
- Real simple syndication/Rich site summary (RSS)
  - "Feeds" provide information about new content

Bibliographic databases

- Contain metadata about (mostly) journal articles and other resources typically found in libraries
- Produced by
  - U.S. government
    - e.g., MEDLINE and subsets, genomics information, etc.
  - Commercial publishers
    - e.g., EMBASE – part of larger SciVal, CINAHL

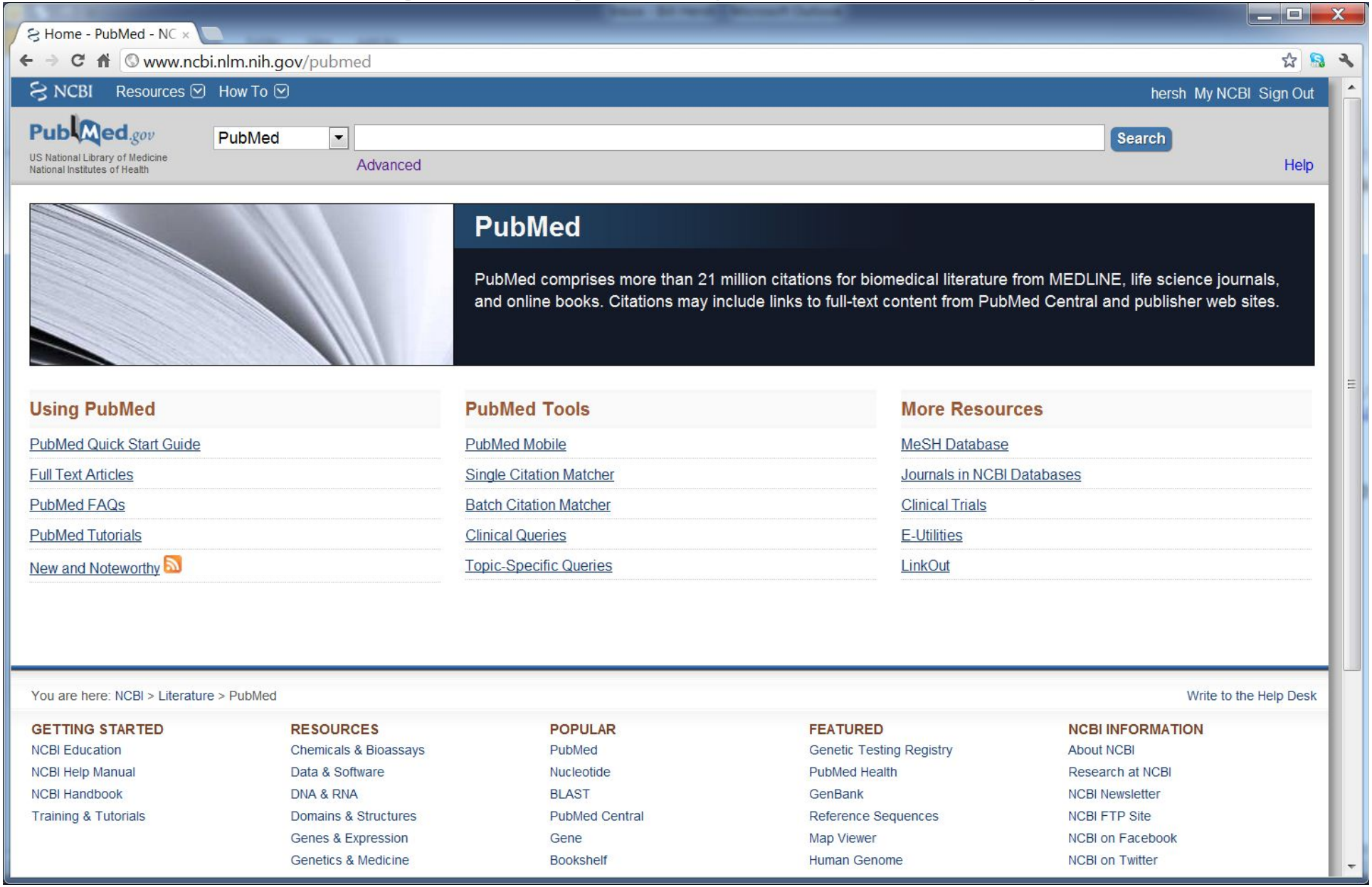
MEDLINE

- References to biomedical journal literature
  - Original medical IR application – launched in 1966
  - Free to world since 1998 via PubMed – pubmed.gov
- Produced by National Library of Medicine (NLM)
- Statistics (http://www.nlm.nih.gov/bsd/bsd_key.html)
  - Over 20 million references to peer-reviewed literature
  - Over 5,000 journals, mostly English language
  - Over 700,000 and growing new references added yearly
- Links to full text of articles and other resources

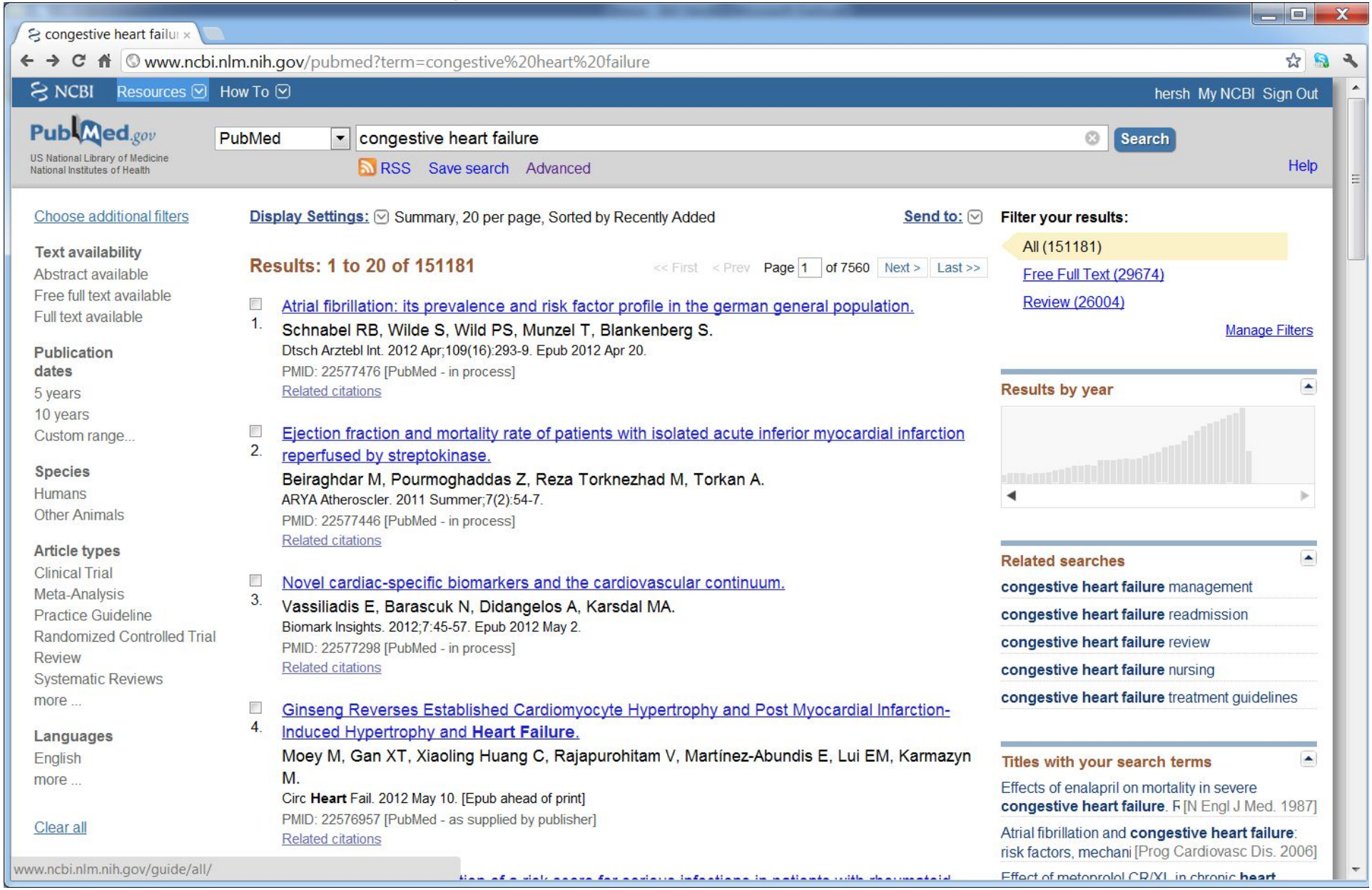
Let's take a tour of PubMed

- User wants to know about treatment of congestive heart failure with angiotensin-converting enzyme (ACE) inhibitors
  - PubMed maps query into appropriate Boolean statement
- Simple AND yields way too many results, so want to narrow down, especially to best evidence
  - Done by applying Limits or using Clinical Queries

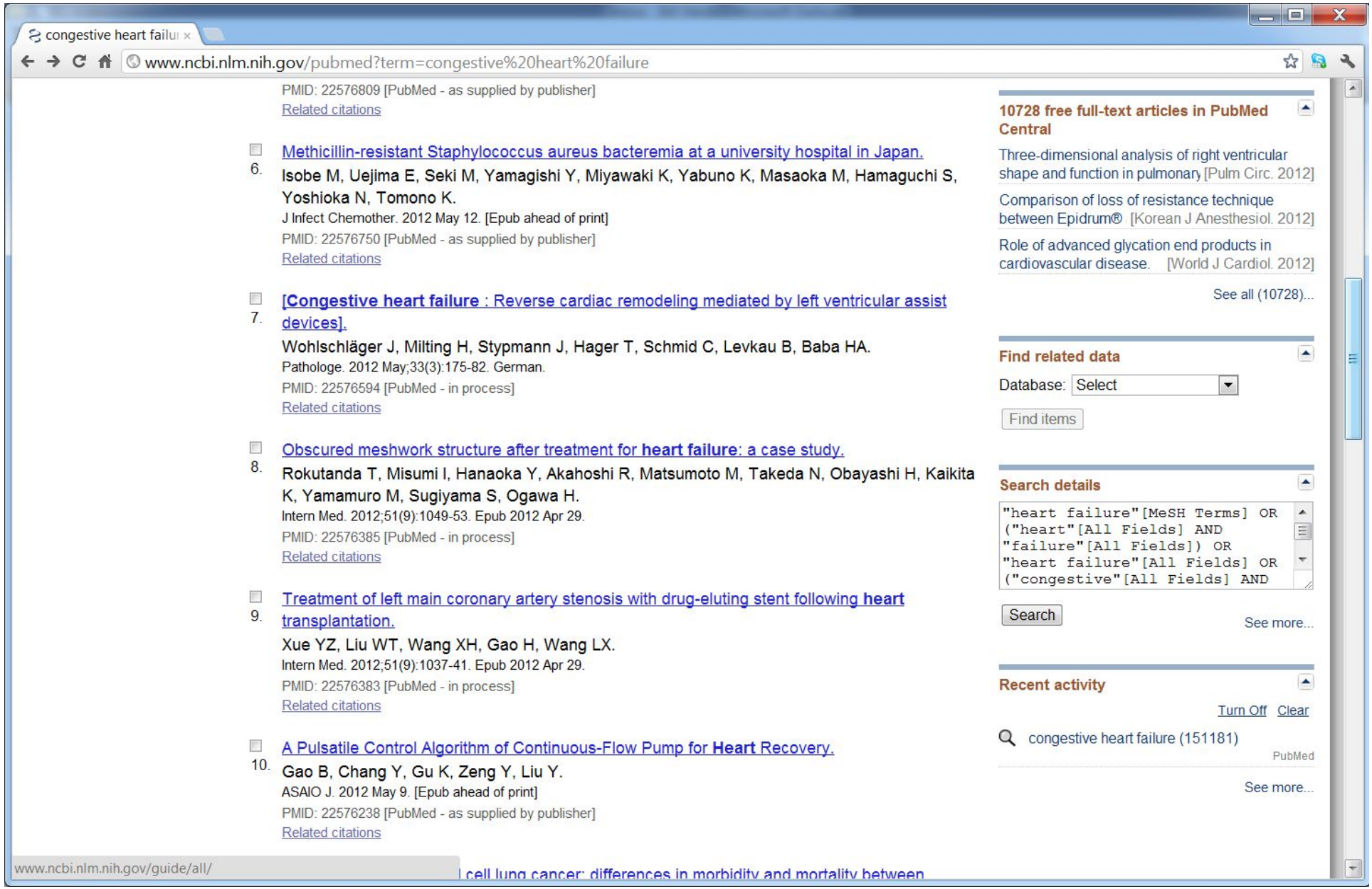
Navigating to pubmed.gov



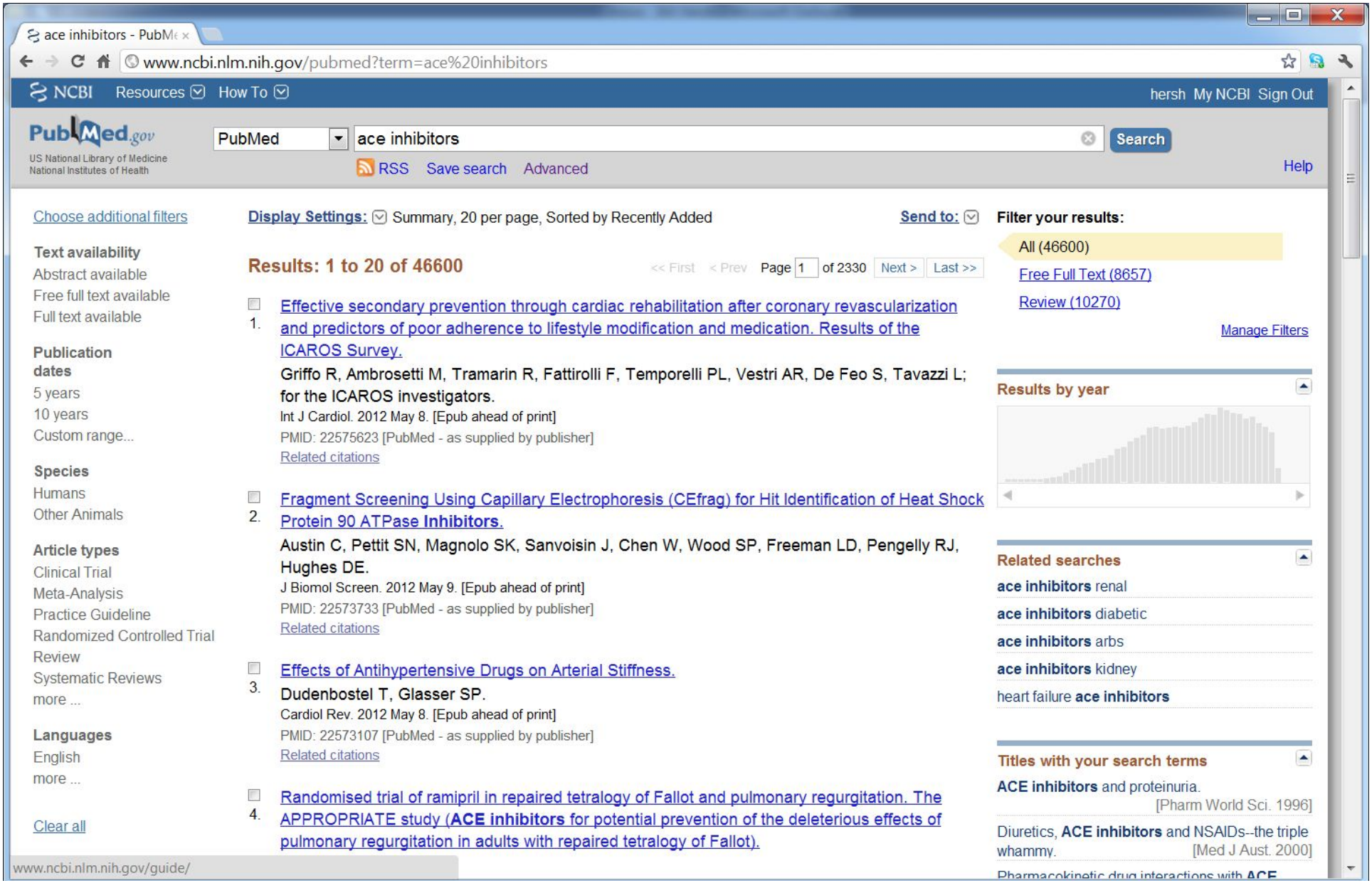
Search on just CHF – note features



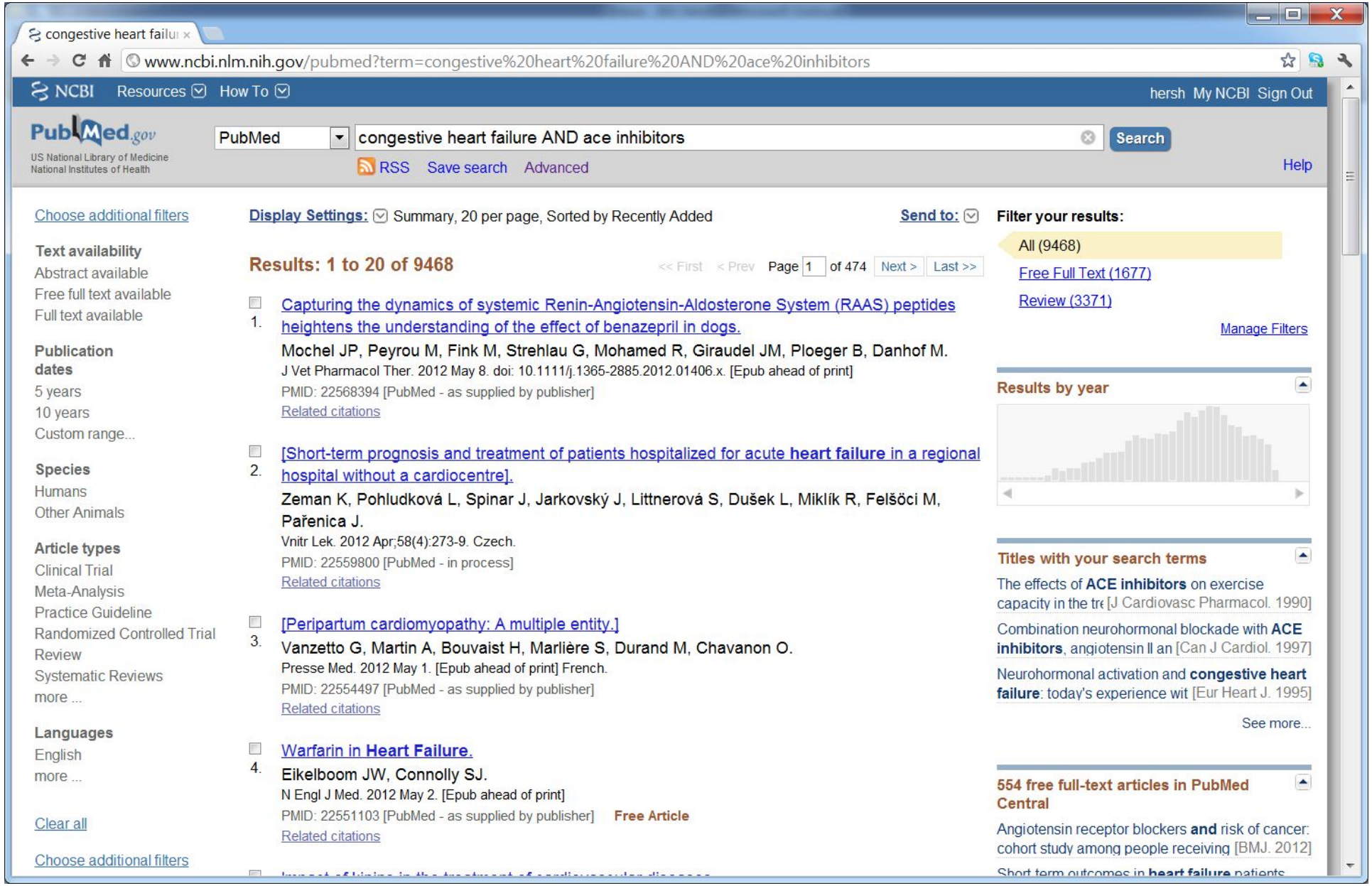
And more features



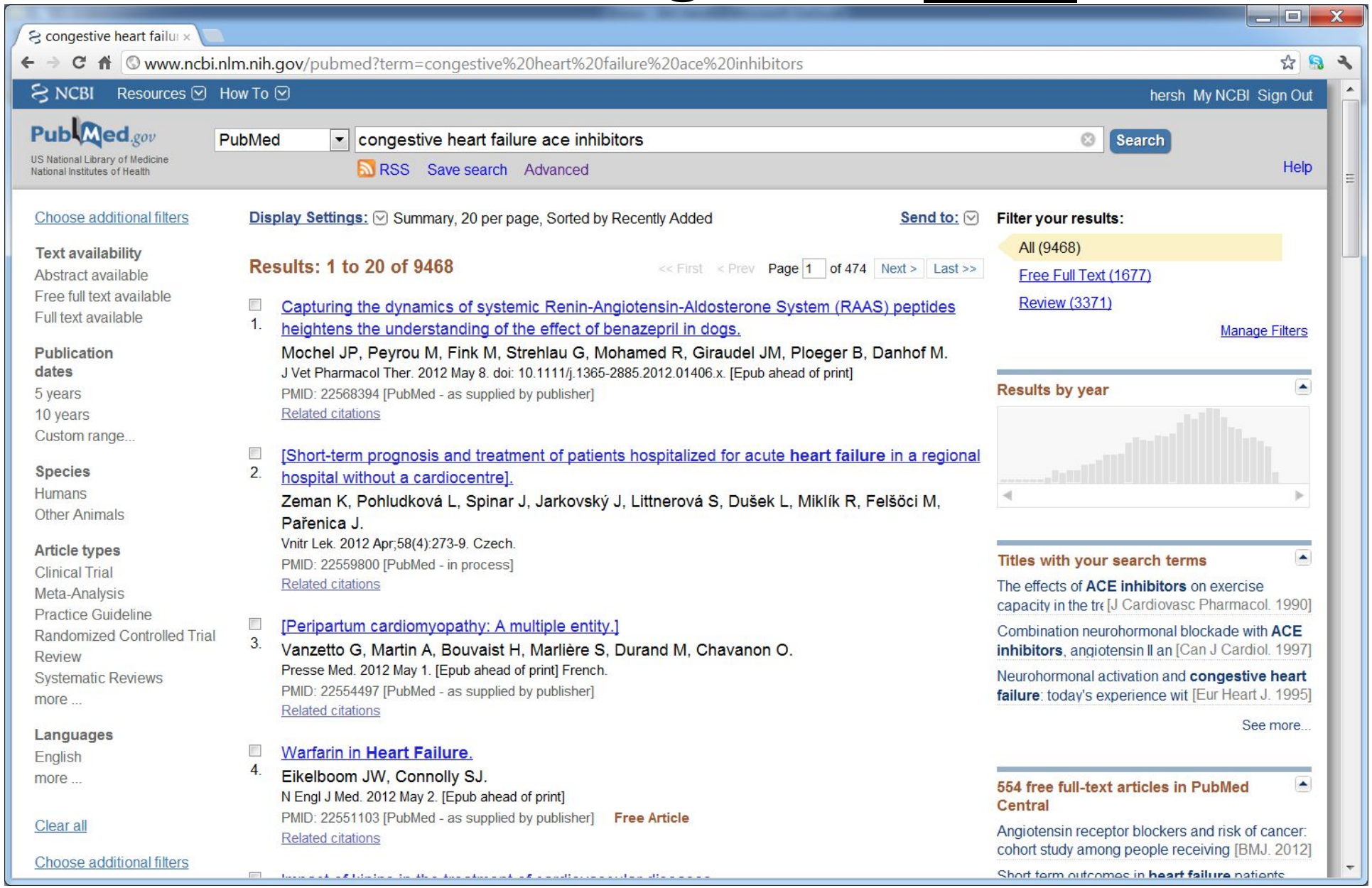
Search on just ACE inhibitors



Need to AND these, but still too many



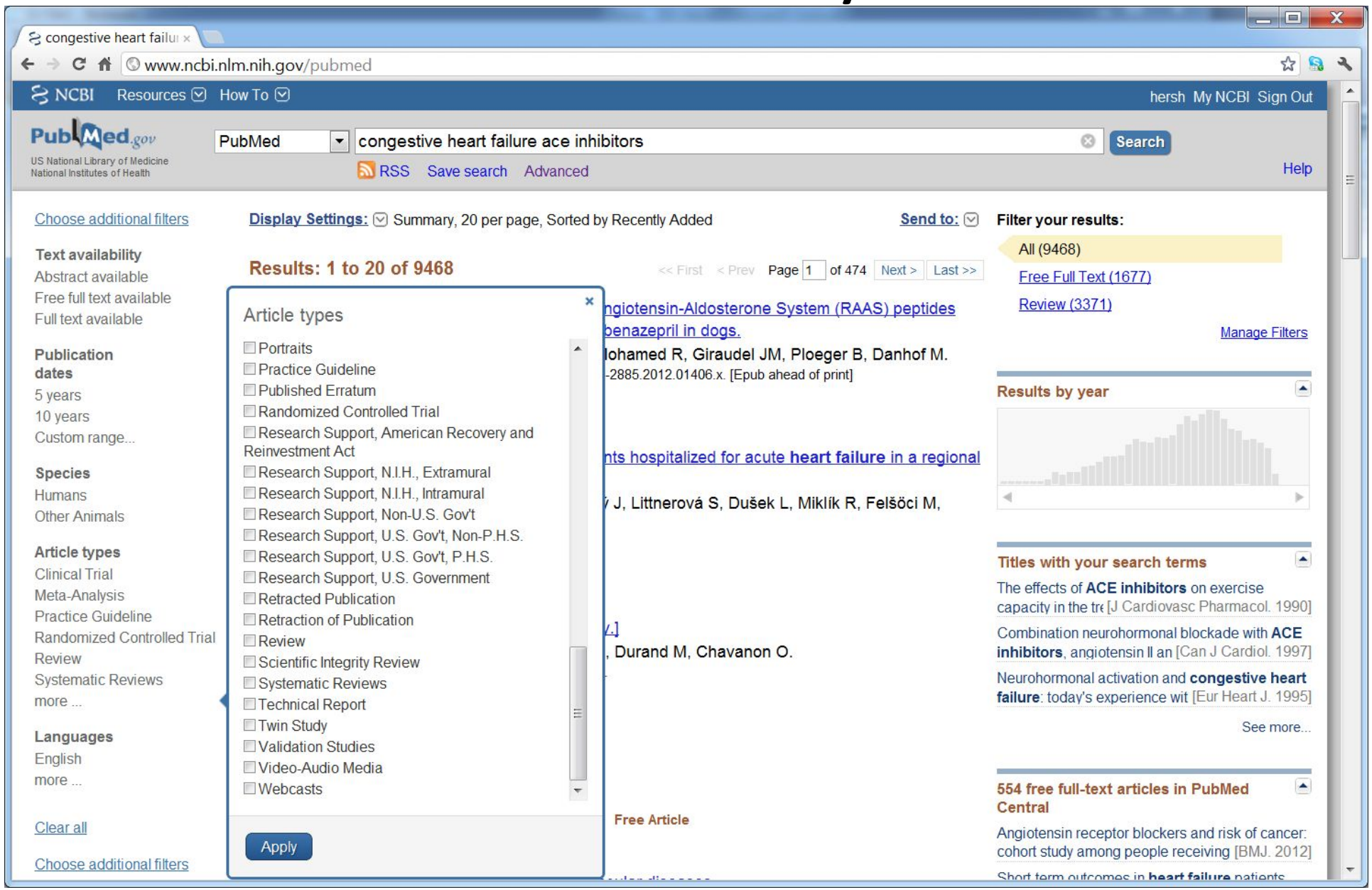
What if I forget the AND?



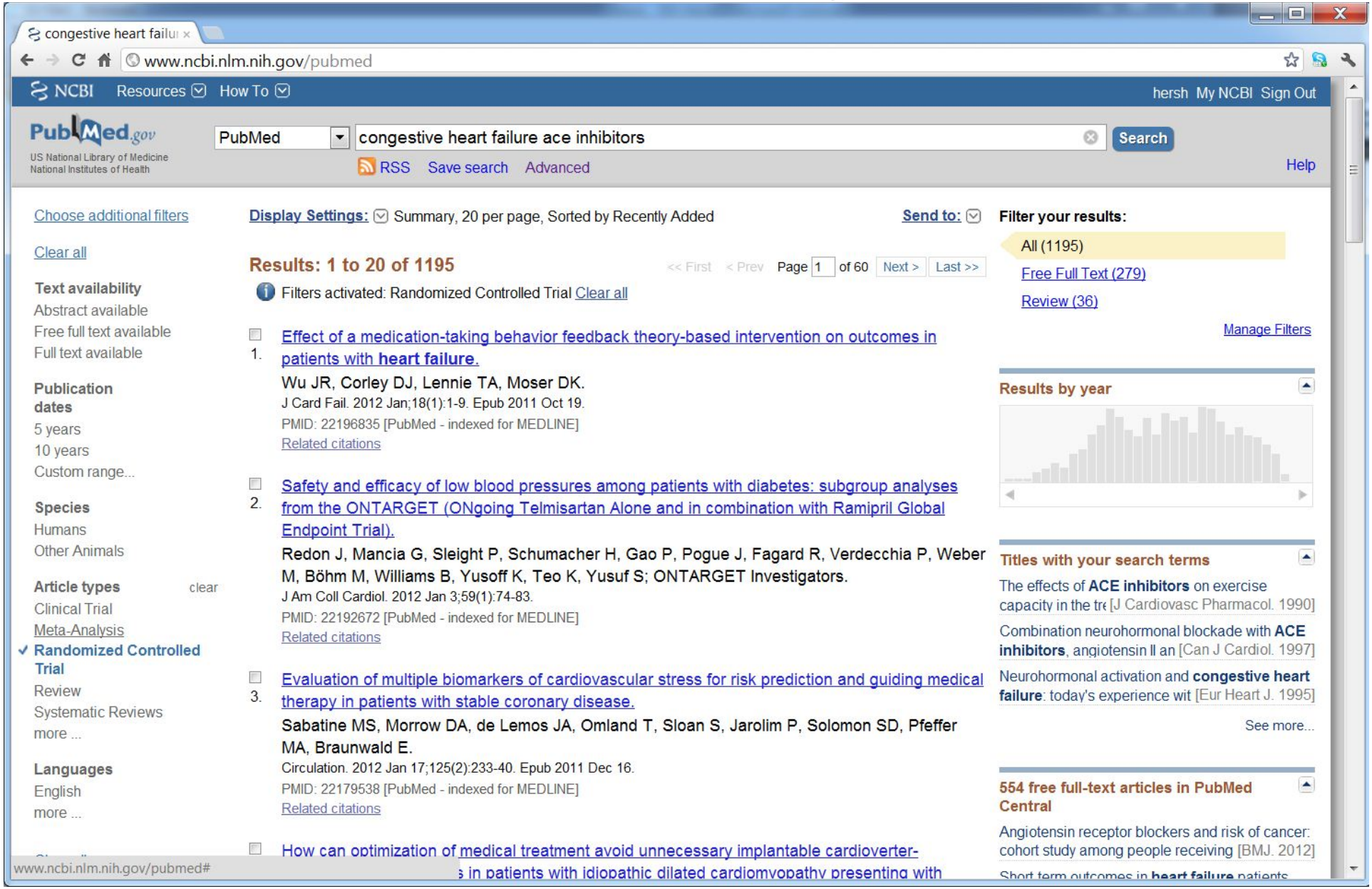
How did it do that?

- PubMed mapping determines terms and appropriate Boolean operators, e.g.,
  - “congestive heart failure ace inhibitors” becomes:
  - ("heart failure"[MeSH Terms] OR ("heart"[All Fields] AND "failure"[All Fields]) OR "heart failure"[All Fields] OR ("congestive"[All Fields] AND "heart"[All Fields] AND "failure"[All Fields]) OR "congestive heart failure"[All Fields]) AND ("angiotensin-converting enzyme inhibitors"[MeSH Terms] OR ("angiotensin-converting"[All Fields] AND "enzyme"[All Fields] AND "inhibitors"[All Fields]) OR "angiotensin-converting enzyme inhibitors"[All Fields] OR ("ace"[All Fields] AND "inhibitors"[All Fields]) OR "ace inhibitors"[All Fields] OR "angiotensin-converting enzyme inhibitors"[Pharmacological Action])

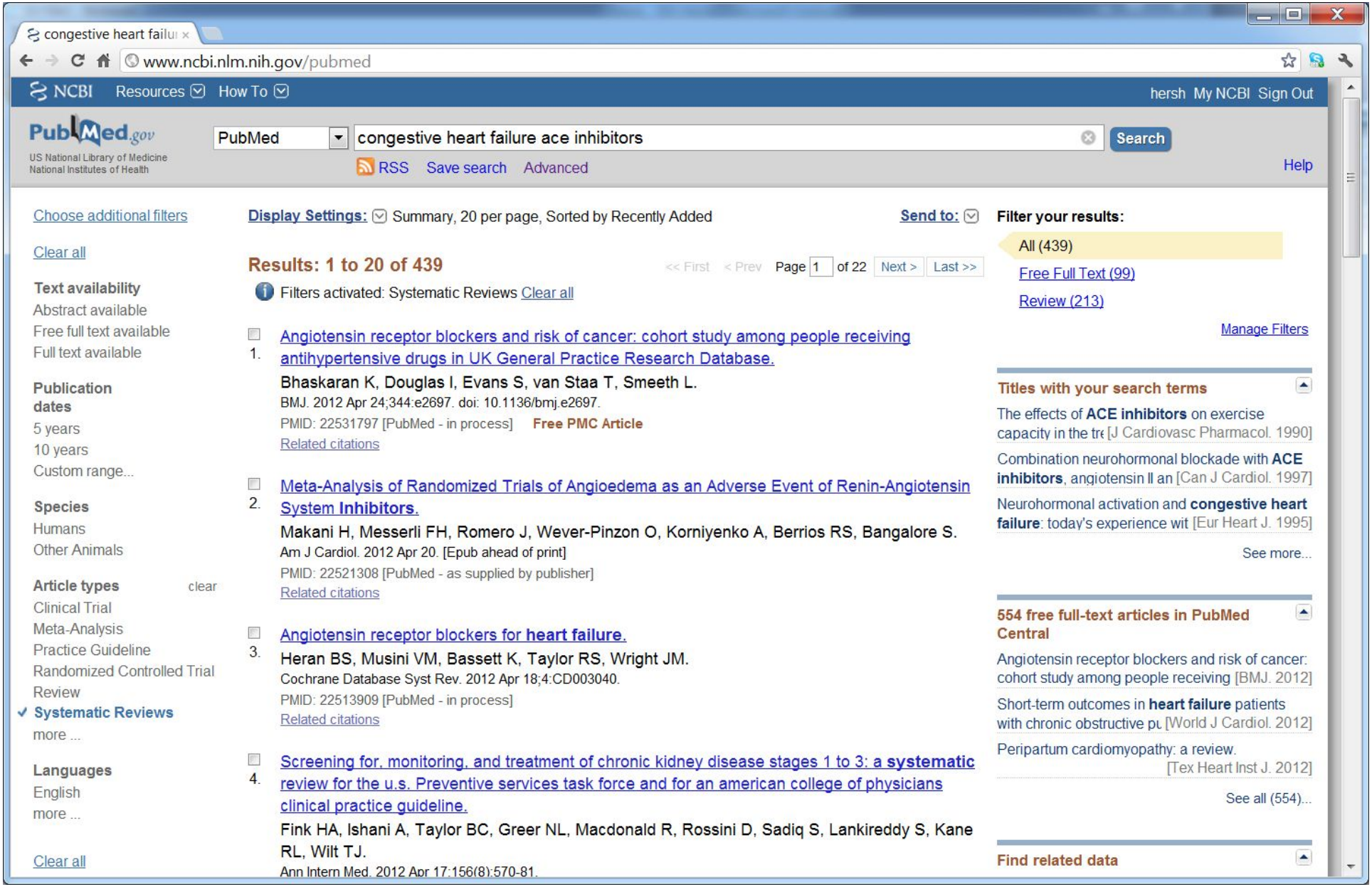
But 9000+ is still way too much



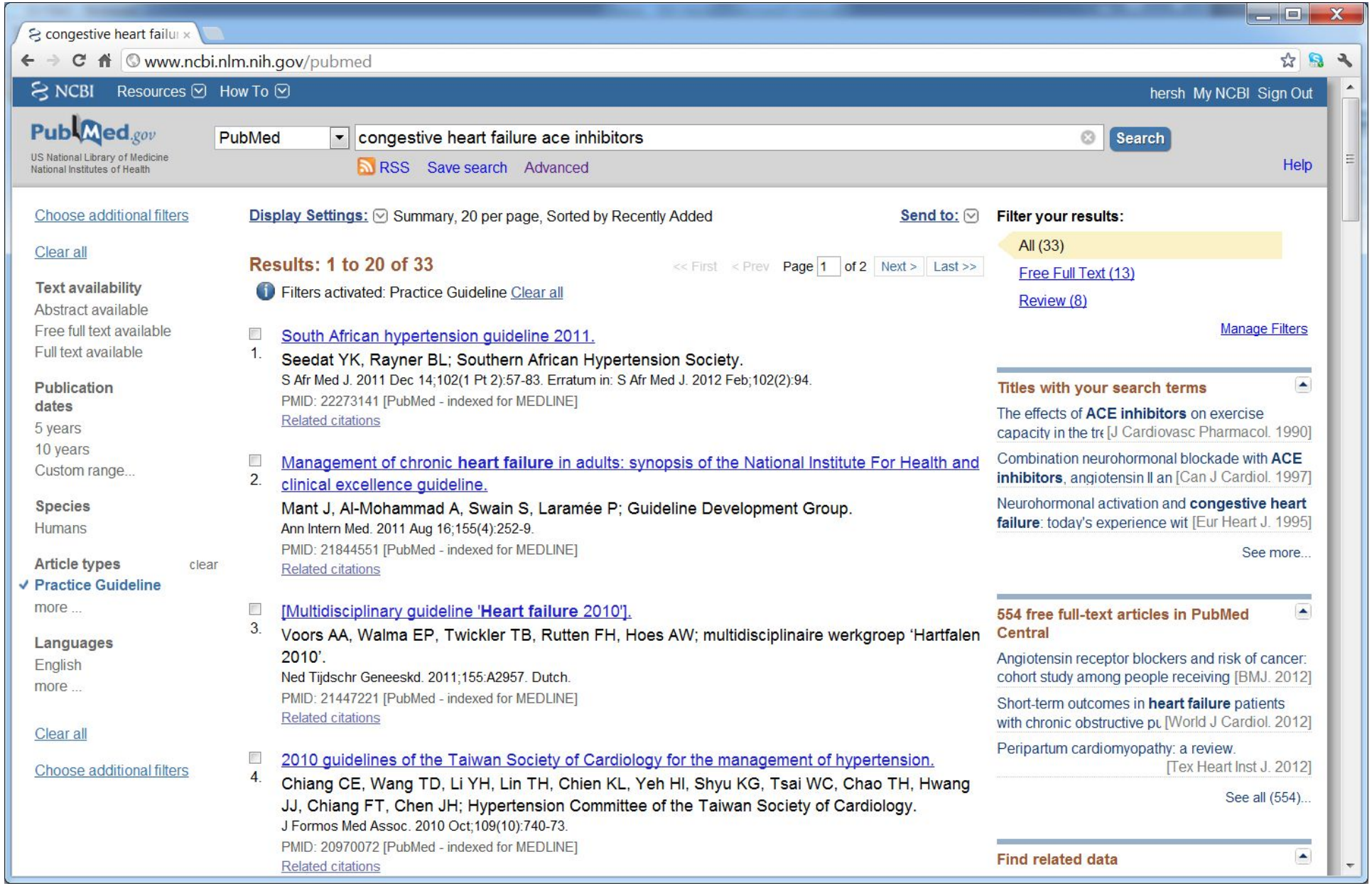
So can limit by RCT



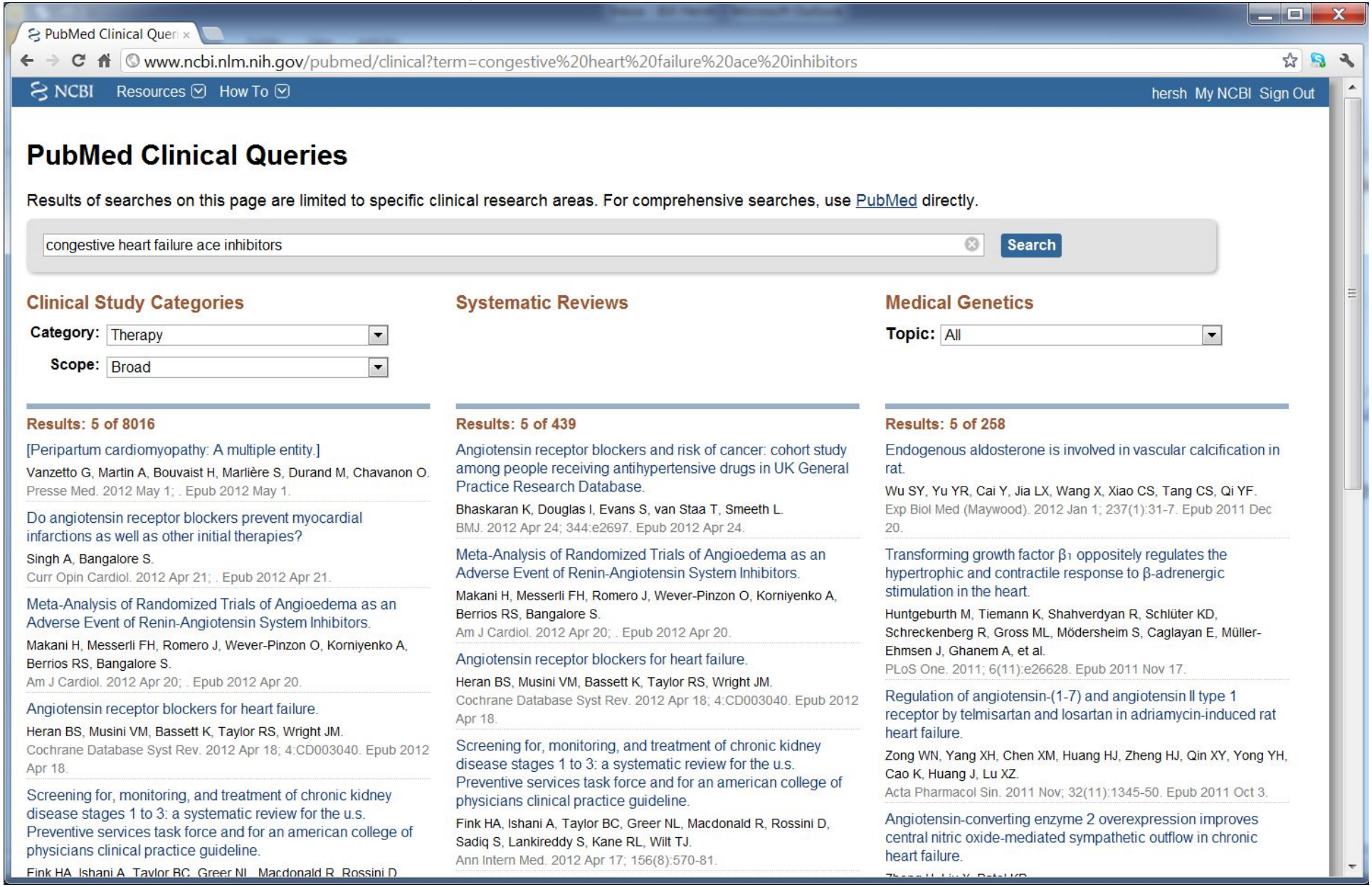
Still too many, so use other limits



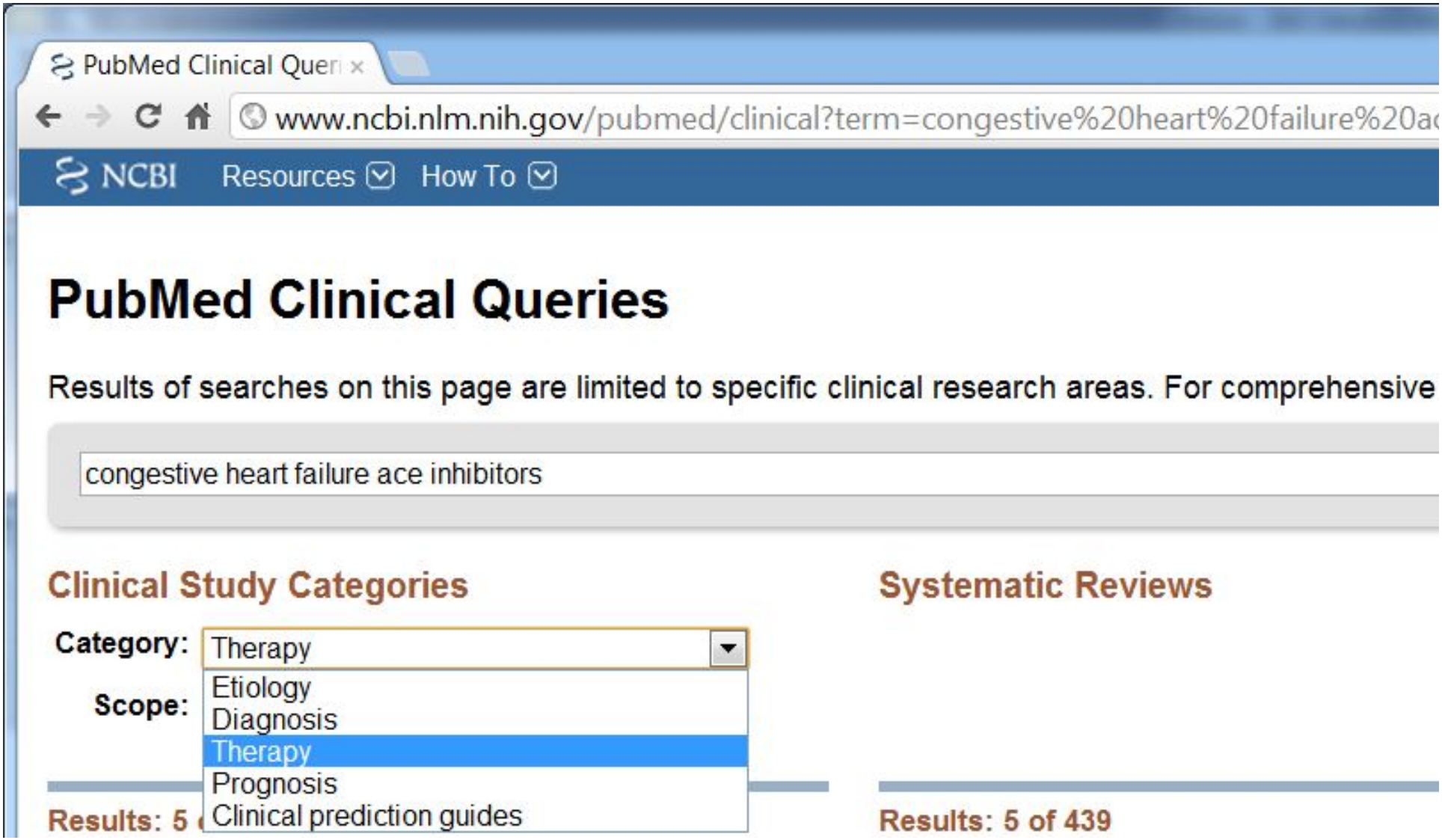
Or further limits



Another option is Clinical Queries



Clinical Queries also allows other question types



National Guidelines Clearinghouse

- Produced by Agency for Healthcare Research and Quality (AHRQ)
  - www.guideline.gov
- Contains detailed information about guidelines
  - Including degree they are evidence-based
  - Interface allows comparison of elements in database for multiple guidelines
- Has links to those that are free on Web and links to producers when proprietary

Web catalogs

- Generally aim to provide quality-filtered Web sites aimed at specific audiences
  - Distinction between catalogs and sites blurry
- Some are aimed towards clinicians
  - HON Select – http://www.hon.ch/HONselect/
  - Translating Research into Practice – www.tripdatabase.com
- Others are aimed towards patients/consumers
  - Healthfinder – www.healthfinder.gov

RSS

- RSS "feeds" provide short summaries, typically of news, journal articles, or other recent postings on Web sites
- Users receive RSS feeds by an RSS aggregator that can typically be configured for the site(s) desired and to filter based on content
  - Work as standalone, in Web browsers, in email clients, etc.
- Two versions (1.0, 2.0) but basically provide
  - Title – name of item
  - Link – URL of full page
  - Description – brief description of page

Full-text content

- Contains complete text as well as tables, figures, images, etc.
- If there is corresponding print version, both are usually identical
- Includes
  - Periodicals
  - Books
  - Web sites – may include either of above

Full-text primary literature

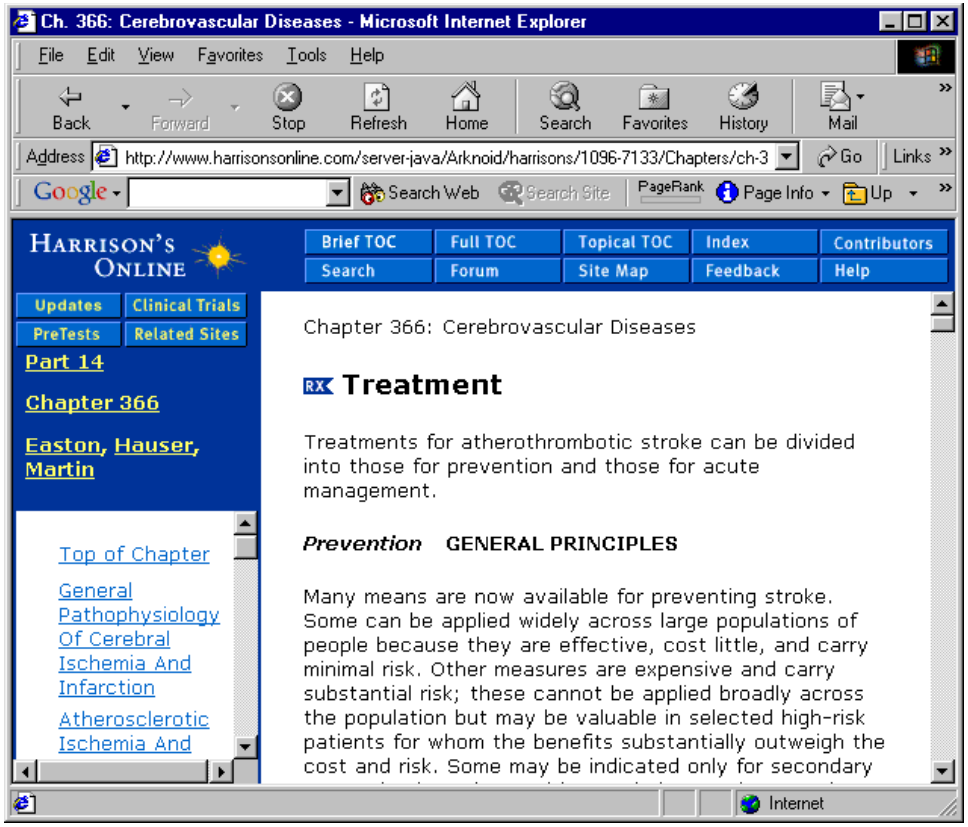
- Almost all biomedical journals available electronically
  - Many published by Highwire Press (www.highwire.org), which adds value to content of original publisher, including *British Medical Journal*, *Journal of the American Medical Association*, *New England Journal of Medicine*, etc.
  - Growing number available via open-access model, e.g., *Biomed Central* (BMC), *Public Library of Science* (PLoS)
- Other publishers license and provide to vendors – e.g., from Ovid, MDConsult, etc.
- Impediments to wider dissemination are economic and not technical (Hersh 2000; McGuigan, 2007)

Books

- Textbooks
  - Most well-known clinical textbooks are now available electronically
    - e.g., *Harrison's Principles of Internal Medicine*
  - Most are bundled into large collections by publishers
    - e.g., Access Medicine, Elsevier, Kluwer (Lippincott Williams & Wilkins)
  - NLM has developed books site as part of PubMed
    - http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books
- Compendia of drugs, diseases, evidence, etc.
- Handbooks – very popular with clinicians

Value added for electronic books

- Multimedia, e.g., skin lesions, shuffling gait of Parkinson's Disease, etc.
- Bundling of multiple books
- Can be updated in "between "editions"
- Linkage to other information, e.g., to references, self-assessments, updates, other resources, etc.



OREGON HEALTH & SCIENCE UNIVERSITY

Web sites

- Plenty of good content can be retrieval by Google, Bing, and other general search engines
  - *Caveat lector et viewor* - let the reader and the viewer beware (Silberg, 1997)
- There are also more narrow coherent collections of information on Web
  - Usually take advantage of Web features, such as linking, multimedia
  - Increasingly integrated with other resources and available on different platforms (e.g., integrated into electronic health records [EHRs], on smartphones, etc.)

Some notable full-text content on Web sites

- Government agencies
  - National Cancer Institute
    - www.cancer.gov
  - Centers for Disease Control – travel and infection information
    - www.cdc.gov
    - http://www.cdc.gov/travel/
  - Other NIH institutes, e.g., National Heart, Lung, and Blood Institute (NHLBI)
    - www.nhlbi.nih.gov

Full-text Web sites (cont.)

- Physician-oriented medical news and overviews, e.g.,
  - Medscape – www.medscape.com
  - PEPID – www.pepid.com
  - Many professional societies provide to members, e.g., http://www.acponline.org/clinical_information/
- Patient/consumer-oriented, e.g.,
  - Intelihealth – www.intelihealth.com
  - NetWellness – www.netwellness.com
  - WebMD – www.webmd.com
- Many mobile apps provide health information, e.g.,
  - iTriage – www.itriagehealth.com

Annotated

- Non-text or structured text annotated with text
- Includes
  - Image collections
  - Citation databases
  - Evidence-based medicine databases
  - Clinical decision support
  - Genomics databases
  - Other databases

Image collections

- Most prominent in the "visual" medical specialties, such as radiology, pathology, and dermatology
- Well-known collections include
  - Visible Human – http://www.nlm.nih.gov/research/visible/visible_human.html
  - BrighamRad – http://brighamrad.harvard.edu/
  - WebPath – http://library.med.utah.edu/WebPath/webpath.html
  - More pathology – PEIR, www.peir.net
  - DermIS – www.dermis.net
  - More dermatology – www.visualdx.com
- Many have associated text, which assists with indexing and retrieval

Citation databases

- *Science Citation Index* and *Social Science Citation Index*
  - Database of journal articles that have been cited by other journal articles
  - Now part of a package called *Web of Science*, which itself is part of a larger product, *Web of Knowledge* (Thomson-Reuters)
    - isiwebofknowledge.com, wokinfo.com
- SCOPUS – http://www.info.sciverse.com/scopus
- Google Scholar – scholar.google.com

Evidence-based medicine databases

- Cochrane Database of Systematic Reviews
  - Collection of systematic reviews, kept updated
- Evidence "formularies"
  - Clinical Evidence – BMJ
  - JAMAevidence
  - PIER (Physician's Information and Education Resource, American College of Physicians) – disease-oriented overviews, tagged for evidence
- Up to Date
  - Clinically oriented overviews of medicine
- InfoPOEMS
  - "Patient-oriented evidence that matters"

Clinical decision support (CDS)

- Content used in CDS systems, usually part of EHRs
  - Order sets (usually "evidence-based")
  - CDS rules
  - Health/disease management templates
- Growing and evolving commercial market for such tools, especially as EHR adoption increases; leaders include
  - Zynx – www.zynxhealth.com
  - Thomson Reuters – thomsonreuters.com
  - EHR vendors themselves and partners

Genomics databases

- National Center for Biotechnology Information (NCBI, www.ncbi.nlm.nih.gov; Sayers, 2012) collection links
  - Literature references – MEDLINE
  - Textbook of genetic diseases – On-Line Mendelian Inheritance in Man
  - Sequence databases – Genbank
  - Structure databases – Molecular Modeling Database
  - Genomes – Catalog of genes
  - Maps – Locations of genes on chromosomes

Other databases

- ClinicalTrials.gov
  - Originally database of clinical trials funded by NIH
  - Now used as register for clinical trials, with results reporting for some (DeAngelis, 2005; Laine, 2007; Zarin, 2011)
- NIH RePORTER
  - http://projectreporter.nih.gov/reporter.cfm
  - Database of all research grants funded by NIH
  - Replaced the CRISP database

Aggregations – integrating many resources

- Clinical: Merck Medicus – www.merckmedicus.com
  - Collection of many resources available to any licensed US physician
- Biomedical research: Model organism databases, e.g., Mouse Genome Informatics
  - www.informatics.jax.org
- Consumer: MEDLINEplus – medlineplus.gov
  - Integrates a variety of licensed resources and public Web sites

Evaluation

- Questions often asked
  - Is system used?
  - Are users satisfied?
  - Do they find relevant information?
  - Do they complete their desired task?
- Most studied group is physicians, with systematic reviews of results (Hersh, 1998, Pluye, 2005)
- Most IR evaluation research has focused on retrieval of relevant documents, which may not capture full spectrum of usage
  - Often consists of challenge evaluations that develop "test collections" – best known is (non-medical) Text Retrieval Conference (TREC, trec.nist.gov) (Voorhees, 2005)

Is system used?

- Most studies done prior to ubiquitous Internet, electronic health records, mobile devices, etc.
- Studies in various clinical settings (Hersh, 2009; Magrabi, 2005) showed average use varied from 0.3 to 8.7 accesses per person-month
- Whatever the actual number, this paled in comparison to known physician information needs (Gorman, 1995) of two questions per every three patients

Are users satisfied?

- Most studies report good user satisfaction, but some interesting studies to note
  - Nielsen (1994) meta-analysis found association (though imperfect) between user satisfaction and ability to use computer systems
  - Most Internet users believe they mostly find information they are seeking (Taylor, 2010; Fox, 2011)

Do they find relevant information?

- Most common approach to evaluation
- Usually measured by relevance-based measures of recall and precision
- With ranked output, can combine recall and precision into aggregate measures
  - Mean average precision (MAP)
  - Binary preference (bpref)
  - Normalized discounted cumulative gain (NCDG)

How well do clinicians search? Early results from Haynes (1990)

Searcher Type	Recall	Precision
Novice clinicians	27%	38%
Expert clinicians	48%	48%
Librarians	49%	57%

Other findings

- Little overlap among retrieval sets
  - Searchers tended to find similar quantities of disparate relevant documents
- Novice searchers satisfied with results
  - Adequate information or ignorant bliss?

Extending evaluation beyond physicians and documents

- Other clinicians
  - Nurses – Rolye, 1995
  - Pharmacists – Wanke, 1988
  - Nurse practitioners – Hersh, 2000; Hersh, 2002
- Biomedical researchers
  - Very little study of their use of IR systems
  - Investigated by TREC Genomics Track (Hersh, 2006; Hersh, 2009) – http://ir.ohsu.edu/genomics/
- Image retrieval – ImageCLEFmed (Hersh, 2006; Hersh, 2009)
  - Retrieval performance related to query type, measure selection
  - http://ir.ohsu.edu/image/

Recall and precision studies yield useful results, but

- Are searchers able to solve their information problems by using system?
  - Some results research have used "task-oriented approach" to measure question-answering
  - Hersh (2002) – use of MEDLINE to answer clinical questions
    - Medical students answered 34% of questions before system, 51% afterwards
    - Nurse practitioner students answered 34% of questions before system but did not change with system
    - Time to answer a question was ~30 minutes
    - No association of recall or precision with correct answering

Another task-oriented study

- Westbrook (2005) – use of online evidence system
  - Physicians answered 37% of questions before system, 50% afterwards
  - Nurse specialists answered 18% of questions before system, 50% afterwards
  - Those who had correct answers had higher confidence in their answers, but those not knowing answer initially had no difference in confidence whether answer right or wrong

How do IR systems impact physician practice? (Pluye, 2004)

- Qualitative study found four themes mentioned by physicians
  - Recall – of forgotten knowledge
  - Learning – new knowledge
  - Confirmation – of existing knowledge
  - Frustration – that system use not successful
- Researchers also noted two additional themes
  - Reassurance – that system is available
  - Practice improvement – of patient-physician relationship

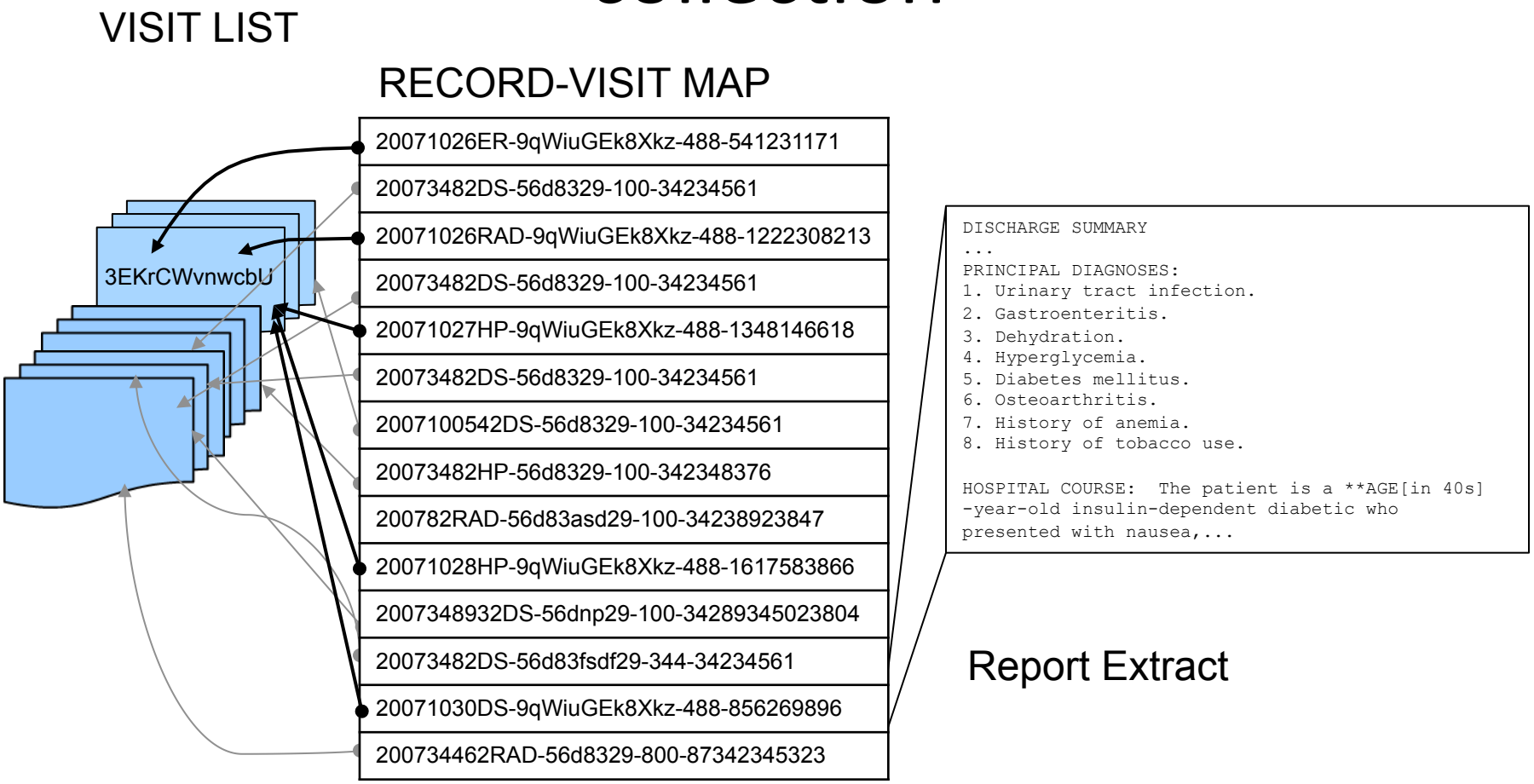
Challenges for IR evaluation moving forward

- Must understand tasks of user and focus evaluation accordingly
- Ultimate measure, like any other informatics application, might be health outcome
  - This may be difficult with IR systems since usage may not directly impact outcomes of patient care or research activity

Research directions – applying IR to medical records

- Most medical records still in narrative documents, where natural language processing (NLP) techniques are improving but still imperfect (Stanfill, 2010)
- For some tasks, can we take an IR approach?
  - TREC Medical Records Track uses de-identified corpus of medical records in initial task of identifying patients as candidates for clinical research studies (Voorhees, 2011)

TREC Medical Records Track test collection



(Courtesy, Ellen Voorhees, NIST)

Topics of TREC Medical Records Track – easy and hard

- Easiest – consistently best results
  - 105: Patients with dementia
  - 132: Patients admitted for surgery of the cervical spine for fusion or discectomy
- Hardest – consistently worst results
  - 108: Patients treated for vascular claudication surgically
  - 124: Patients who present to the hospital with episodes of acute loss of vision secondary to glaucoma
- Large differences between best and worst results
  - 125: Patients co-infected with Hepatitis C and HIV
  - 103: Hospitalized patients treated for methicillin-resistant Staphylococcus aureus (MRSA) endocarditis
  - 111: Patients with chronic back pain who receive an intraspinal pain-medicine pump

Another research direction: question-answering

- Users may retrieve documents, but usually want answers to questions
- Subarea of IR research has focused on question-answering systems (Strzalkowski, 2006)
- Most recent "hype" of question-answering is the IBM Watson system
  - Developed out of TREC Question-Answering Track (Voorhees, 2005; Ferrucci, 2010)
  - Beat humans at Jeopardy (Markoff, 2011)
  - Now being applied to healthcare (Ferrucci, 2012)

Conclusions

- Mine
  - IR, including in biomedicine, has become widespread and mainstream
  - Challenges still exist, especially in specific domains and/or for specific tasks
  - Plenty of room left for research but building on top of existing systems and not de novo
- Yours?