# Functional Search Problems Among MEDLINE Databases

Sean Burns
in collaboration with Tyler Nix (U-M), Robert Shapiro, and Jeffrey Huber
~~March 25, 2020~~ September 9, 2020

Or, how to derail your research plans (and threaten tenure likelihood): from scholarly communication to information retrieval

# Purpose of the Study

- Since MEDLINE data is provided by multiple information service providers, and users access and use these different versions of MEDLINE, that are all based on the same data file, the goal of this study is to identify where and how MEDLINE-based bibliographic database platforms differ when searched with semantically and logically equivalent search queries/strategies

# Motivation of the Study

- Bibliographic databases enable literature searches, which is a major part of all science and scholarship in general
- MEDLINE is the most important bibliographic database in the health, medical, and biosciences and fundamental to research like systematic reviews (SR)
- Systematic review research is the primary method in gathering evidence to support clinical interventions and to reduce "bias by identifying, appraising, and synthesizing all relevant studies on a particular topic"
  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3024725/
- Yet up to 60% of SRs may "not retrieve 95% of all available relevant references as many fail to search important databases" https://doi.org/10.1186/s13643-017-0644-y

Originally, the plan was to study megajournals

# Research on Megajournals

- Megajournals (ex. PLOS ONE) are open access born-digital journals that
- More fully embrace online publishing than born-print journals.
- In practice, this means:
  - Optional volumes and/or issues (as-ready publishing--daily)
    - In an online format, this affects how users (e.g., readers) interact with the site
      - Search engine traffic becomes more important
      - Digital library frameworks become more important
        - Table of contents are replaced with *collections* and *exhibits* by topic or by theme
  - Disciplinary boundaries are relaxed
    - Authors cite a wider range of journal titles when examining the journal as a whole (this influences search traffic and web search rankings)

# Research on Megajournals

- Megajournals (ex. PLOS ONE) are open access born-digital journals that
- More fully embrace online publishing than born-print journals.
- In practice, this means:
  - No constraints on printed pages per year (journal costs are not tied to alloted pages per year by a publisher)
    - Increases available supply (can publish more papers)
    - Increases the acceptance rate (no longer a measure of selectivity)
    - Makes the Journal Impact Factor (JIF/IF), historically and problematically a proxy for journal quality, meaningless
      - JIF assumes a fairly constant publication rate per journal per year. When the # of publications is held fairly constant then the JIF measures average annual changes in citation counts; with megajournals, the # of publications is not a constant

From all citing sources, inc. Journal X, in the bibliographic database

JIF = Citations in 2019 to

Number of citable items in 2017 + 2018

Ex: research articles, reviews published in Journal X

# JAMA, JIF 2009-2019

Relatively Stable Publishing History (-7.5% overall drop in published, citable items)

Cf. to 46% increase in overall citations, 2009-2019

| Year | Citable Items | Citations Year | 2 Year JIF |
|---|---|---|---|
| 2009 | 454 | 13120 | 28.9 |
| 2010 | 459 | 13775 | 30.0 |
| 2011 | 467 | 14022 | 30.0 |
| 2012 | 453 | 13580 | 30.0 |
| 2013 | 452 | 13735 | 30.4 |
| 2014 | 453 | 15986 | 35.3 |
| 2015 | 449 | 16920 | 37.7 |
| 2016 | 425 | 18872 | 44.4 |
| 2017 | 410 | 19541 | 47.7 |
| 2018 | 421 | 21586 | 51.3 |
| 2019 | 420 | 19127 | 45.5 |

~16pt overall JIF increase since citations increased but citable items (denominator) remained 'kinda' stable

Source: *Web of Science* InCites Journal Citation Reports

# PLOS ONE, JIF 2009-2019

Overall 868% increase in published, citable items

Cf. 510% overall increase in citations

| Year | Citable Items | Citations Year | 2 Year JIF |
|---|---|---|---|
| 2009 | 3954 | 17204 | 4.4 |
| 2010 | 7120 | 31404 | 4.4 |
| 2011 | 11125 | 45521 | 4.1 |
| 2012 | 20503 | 76475 | 3.7 |
| 2013 | 37229 | 131563 | 3.5 |
| 2014 | 54945 | 177706 | 3.2 |
| 2015 | 61541 | 188116 | 3.1 |
| 2016 | 58157 | 163193 | 2.8 |
| 2017 | 50188 | 138835 | 2.8 |
| 2018 | 42458 | 117863 | 2.8 |
| 2019 | 38271 | 104864 | 2.7 |

-1.7pt overall drop because the number of citatable items (denominator) is not stable though citations increased

Source: *Web of Science* InCites Journal Citation Reports

# Megajournal Study

Published one of the first studies on megajournals, a new one at the time called *PeerJ*

Burns, C.S. (2015). Characteristics of a megajournal: A bibliometric case study. *Journal of Information Science Theory and Practice, 3*(2), 16-30. doi:https://doi.org/10.1633/JISTaP.2015.3.2.2 data doi:http://doi.org/10.6084/m9.figshare.1501498 code: https://github.com/cseanburns/peerj

The above paper started to accumulate a few citations as others started researching megajournals. I was excited that I found a new research space! On to tenure!

# But then, Derailment...

In preparation for a follow up megajournal study ...

- Noticed a bibliographic database discrepancy because of what seemed to be an information retrieval issue
- Since I was looking at medical journals, I went to @Huber and @Shapiro to consult, and recruited @Nix later
- They shared separate observations they saw when searching medical databases
- The issues we discussed were highly problematic and very basic to librarianship and information {science,retrieval}

# Library and Information Science > Information Retrieval

Some basics first.

# Library and Information Science (LIS)

- Is based on a core principle of **information access**
- Pertinent to this talk, **information access** is studied and practiced under two (among other) concepts:
  - **Information organization** (aka, **knowledge organization**)
  - **Information retrieval**

# Information Organization (metadata)

- Descriptive organization (or cataloging, indexing, etc.); information taken from the information package
  - Title of information package
  - Creator/author/contributor/etc of information package
  - Publication date of information package, etc
- Subject (aspect of DO) organization (or cataloging, indexing, etc.); information inferred from the information in the package
  - The **aboutness** of an information package
    - Most theoretical and empirical work in {information,knowledge} organization is done here
    - Generally applied in two ways:
      - As part of a heading or thesauri system (LCSH, MeSH, etc)
      - As part of a classification system (LCC, Dewey, etc)
      - Community driven (folksonomies, {hash,}tags, etc)

# Subject Organization

Generally undertaken in two ways. One:

- Manually (e.g., a librarian): someone does a subject analysis of a work, based on literary warrant (or user, epistemic, ethical, gender, ... warrant), and derives **index** terms to organize and describe the work:
  - "Catalogers are not only transcribers, they are seekers of truth."
  - Pierson, Harriet Wheeler. (1934). The Forest of Pencils: Adventures in Corporate Entry. *The Library Quarterly, 4*(2), 306-313. url:http://www.jstor.org/stable/4302077

# Subject Organization

Generally undertaken in two ways. Two:

- Algorithmically (e.g., Google): someone designs an algorithm that derives **index** terms from text in order to organize and describe the work, e.g., identifying frequency of terms in a document and then weighting those terms by their frequency in a document collection, or *corpus*: *tf*idf:*
  - Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*(1), 11–21. https://doi.org/10.1108/eb026526
  - Sparck Jones, K. (1973). Index term weighting. *Information Storage and Retrieval, 9*(11), 619–633. https://doi.org/10.1016/0020-0271(73)90043-0

# Information Retrieval & Information Organization

- Information retrieval is "the selective, systematic recall of logically stored information" (aka, organized information) (Cleveland & Cleveland, 1983, 33).
  - Regardless whether the information is organized manually or algorithmically
  - Cleveland, D. B., & Cleveland, A. D. (1983). *Introduction to indexing and abstracting.* Libraries Unlimited, Ltd.
- Other things are involved, too, like usability, information need, psychological relevance, etc, but
- For our purposes, we're focused on information retrieval as a function of information organization.

# MEDLINE Study

- What is a bibliographic database?
- What is MEDLINE?
- What is MeSH (Medical Subject Headings)?

# What is a Bibliographic Database?

- A bibliographic database is a database that searches and returns bibliographic records (not necessarily full text).
- Bibliographic records include information (metadata) that describes resources in a collection.
  - Descriptive information (metadata)
    - Subject (aboutness) information (metadata)

# What is MEDLINE?

- "MEDLINE is the U.S. National Library of Medicine® (NLM) premier bibliographic database that contains more than 26 million references to journal articles in life sciences with a concentration on biomedicine. A distinctive feature of MEDLINE is that the records are indexed with NLM Medical Subject Headings (MeSH®)." https://www.nlm.nih.gov/bsd/medline.html
- Bibliographic records in MEDLINE contain all the common descriptive information found in bibliographic databases but also MeSH descriptors from the MeSH thesaurus (subject description/metadata).
- A thesaurus is a type of controlled vocabulary that includes a hierarchical set of terms that may have broader or narrower relations to other terms.

# What is MEDLINE?

- MEDLINE is searched online through PubMed: https://pubmed.ncbi.nlm.nih.gov/
- MEDLINE bibliographic records are a large subset of PubMed records
- The National Library of Medicine (NLM) licenses MEDLINE records to other information service providers including:
  - EBSCOhost
  - Ovid
  - ProQuest
  - Web of Science
- That means there are at least four additional platforms to search MEDLINE, each with a different interface, search fields, etc

# MeSH: Organized Biomedical Information

MeSH **descriptors** (or **terms**) are organized into 16 categories (**branches**) and then **subcategories**: from general descriptors to specific descriptors. Updated annually, there are a total of 29,351 descriptors (terms, subject headings) altogether as of 2019.

Each **descriptor** (or **term**) is also called a **subject heading** or just **heading** (the terminology can get confusing).

*descriptor = term = subject heading*

# MeSH: Main Branches

A. Anatomy
B. Organisms
C. Diseases
D. Chemicals and Drugs
E. Analytical, Diagnostic and Therapeutic Techniques and Equipment
F. Psychiatry and Psychology
G. Phenomena and Processes
H. Disciplines and Occupations
I. Anthropology, Education, Sociology and Social Phenomena
J. Technology, Industry, Agriculture
K. Humanities
L. Information Science
M. Named Groups
N. Health Care
V. Publication Characteristics
Z. Geographicals

Source: https://www.nlm.nih.gov/bsd/disted/meshtutorial/meshtreestructures/index.html

# MeSH: Qualifiers / Subheadings

Each **subject heading** can have a **qualifier** (also called a **subheading**), an additional term applied in a bibliographic record that adds more specificity but that is not on the MeSH tree. Thus, qualifiers are floating terms and can be used anywhere applicable on the entire MeSH tree. There are currently 76 of qualifiers.
https://www.nlm.nih.gov/bsd/indexing/training/SUB_010.html

*qualifier = subheading*

# MeSH: Qualifiers / Subheadings

## MeSH Qualifiers with Scope Notes

**Listed below are MeSH Topical Qualifiers listed by Name, Abbreviation, and Short Form. Each Qualifier is defined by a Scope Note that provides guidance on how it should be used.**

**abnormalities - AB, - ABNORM**
Used with organs for congenital defects producing changes in the morphology of the organ. It is used also for abnormalities in animals.

**administration & dosage - AD, - ADMIN**
Used with drugs for dosage forms, routes of administration, frequency and duration of administration, quantity of medication, and the effects of these factors.

**adverse effects - AE, - ADV EFF**
Used with drugs, chemicals, or biological agents in accepted dosage - or with physical agents or manufactured products in normal usage - when intended for diagnostic, therapeutic, prophylactic, or anesthetic purposes. It is used also for adverse effects or complications of diagnostic, therapeutic, prophylactic, anesthetic, surgical, or other procedures.

**agonists - AG, - AGON**
Used with chemicals, drugs, and endogenous substances to indicate substances or agents that have affinity for a receptor and intrinsic activity at that receptor. (From Textbook of Pharmacology, 1991, p.16)

**analogs & derivatives - AA, - ANALOGS**
Used with drugs and chemicals for substances that share the same parent molecule or have similar electronic structure but differ by the addition or substitution of other atoms or molecules. It is used when the specific chemical heading is not available and no appropriate group heading exists.

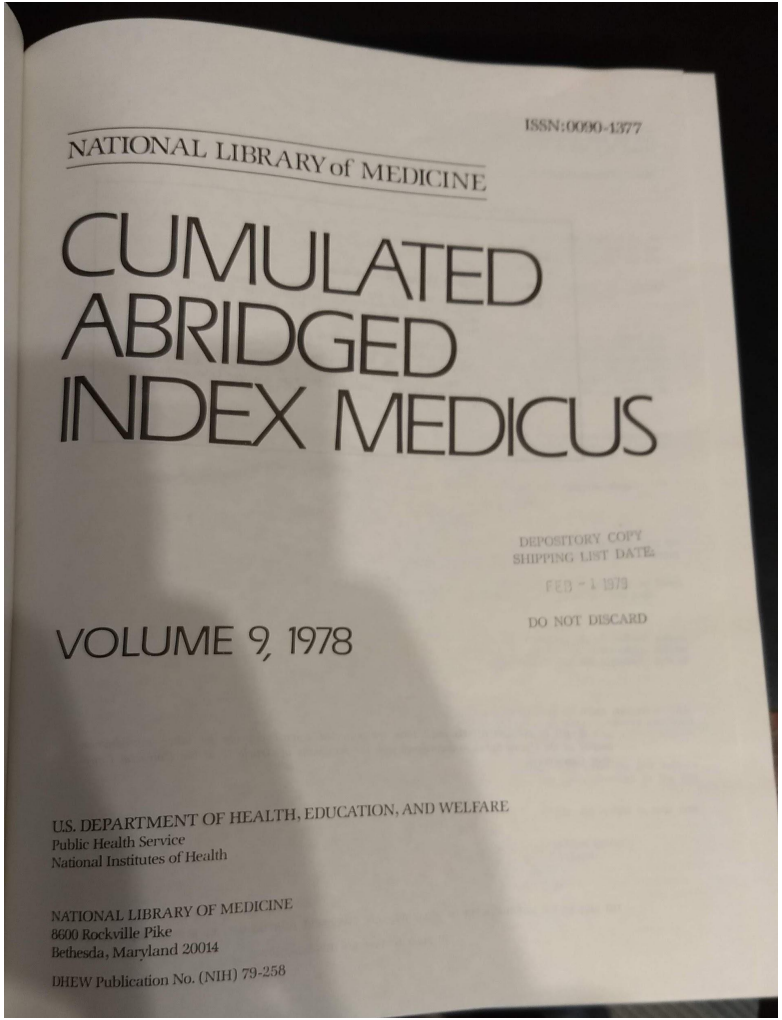
**analysis - AN, - ANAL**
Used for the identification or quantitative determination of a substance or its constituents and metabolites; includes the analysis of air, water, or other environmental carrier. It excludes the chemical analysis of tissues, tumors, body fluids, organisms, and plants for which "chemistry" is used. The concept applies to both methodology and results. For analysis of substances in blood, cerebrospinal fluid, and urine the specific subheading designating the fluid is used.

**anatomy & histology - AH, - ANAT**
Used with organs, regions, and tissues for normal descriptive anatomy and histology, and for the normal anatomy and structure of animals and plants.

**antagonists & inhibitors - AI, - ANTAG**
Used with chemicals, drugs, and endogenous substances to indicate substances or agents which counteract their biological effects by any mechanism.

**biosynthesis - BI, - BIOSYN**
Used for the anabolic formation of chemical substances in organisms, in living cells, or by subcellular fractions.

Source:
https://wayback.archive-it.org/org-350/20191102205211/https://www.nlm.nih.gov/mesh/topsubscope.html




# In the olden days

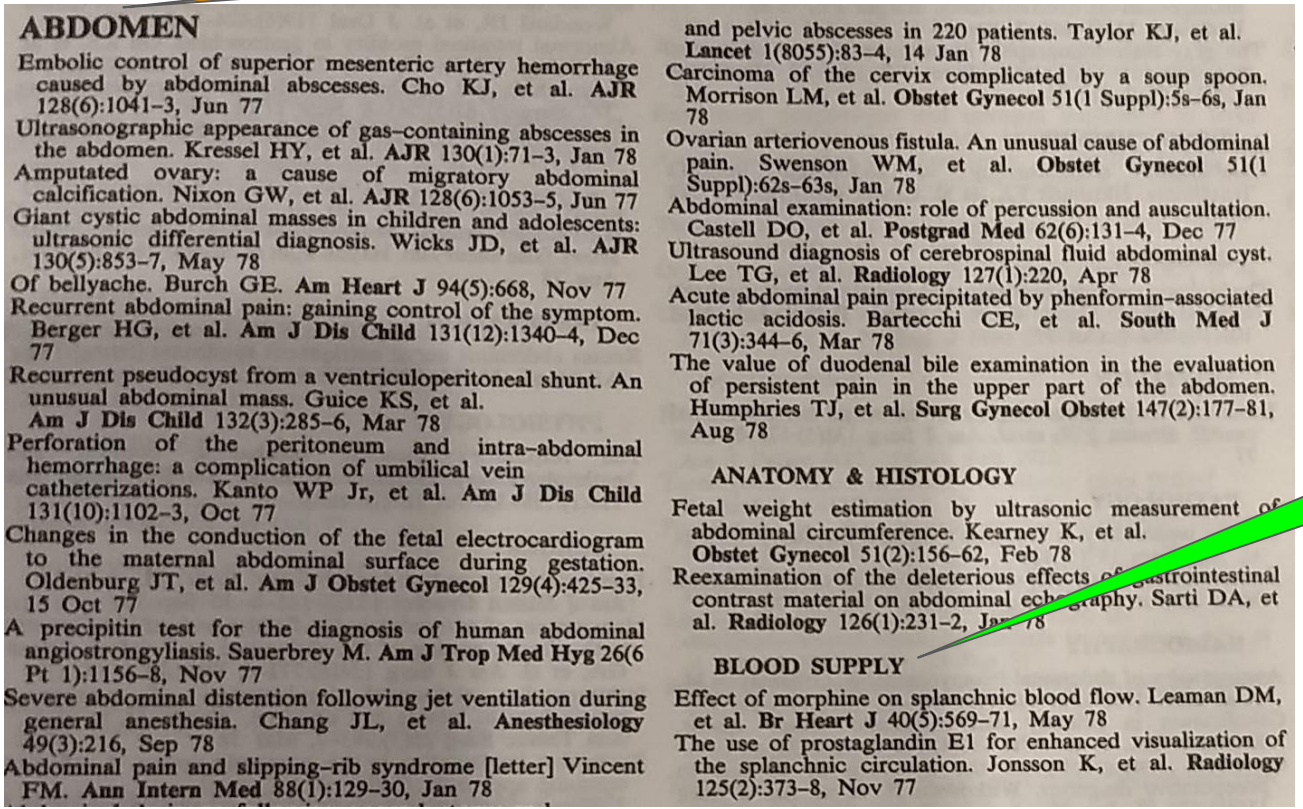
Searching for medical literature in the olden days.

Source: Volume 9, 1978 *Cumulated Abridged Index Medicus*, photo taken at UK Libraries, 02/20/2020.

# In the olden days

Descriptor, Heading, etc.

Qualifier / Subheading



## ABDOMEN

Embolic control of superior mesenteric artery hemorrhage caused by abdominal abscesses. Cho KJ, et al. **AJR** 128(6):1041–3, Jun 77

Ultrasonographic appearance of gas–containing abscesses in the abdomen. Kressel HY, et al. **AJR** 130(1):71–3, Jan 78

Amputated ovary: a cause of migratory abdominal calcification. Nixon GW, et al. **AJR** 128(6):1053–5, Jun 77

Giant cystic abdominal masses in children and adolescents: ultrasonic differential diagnosis. Wicks JD, et al. **AJR** 130(5):853–7, May 78

Of bellyache. Burch GE. **Am Heart J** 94(5):668, Nov 77

Recurrent abdominal pain: gaining control of the symptom. Berger HG, et al. **Am J Dis Child** 131(12):1340–4, Dec 77

Recurrent pseudocyst from a ventriculoperitoneal shunt. An unusual abdominal mass. Guice KS, et al. **Am J Dis Child** 132(3):285–6, Mar 78

Perforation of the peritoneum and intra–abdominal hemorrhage: a complication of umbilical vein catheterizations. Kanto WP Jr, et al. **Am J Dis Child** 131(10):1102–3, Oct 77

Changes in the conduction of the fetal electrocardiogram to the maternal abdominal surface during gestation. Oldenburg JT, et al. **Am J Obstet Gynecol** 129(4):425–33, 15 Oct 77

A precipitin test for the diagnosis of human abdominal angiostrongyliasis. Sauerbrey M. **Am J Trop Med Hyg** 26(6 Pt 1):1156–8, Nov 77

Severe abdominal distention following jet ventilation during general anesthesia. Chang JL, et al. **Anesthesiology** 49(3):216, Sep 78

Abdominal pain and slipping–rib syndrome [letter] Vincent FM. **Ann Intern Med** 88(1):129–30, Jan 78

and pelvic abscesses in 220 patients. Taylor KJ, et al. **Lancet** 1(8055):83–4, 14 Jan 78

Carcinoma of the cervix complicated by a soup spoon. Morrison LM, et al. **Obstet Gynecol** 51(1 Suppl):5s–6s, Jan 78

Ovarian arteriovenous fistula. An unusual cause of abdominal pain. Swenson WM, et al. **Obstet Gynecol** 51(1 Suppl):62s–63s, Jan 78

Abdominal examination: role of percussion and auscultation. Castell DO, et al. **Postgrad Med** 62(6):131–4, Dec 77

Ultrasound diagnosis of cerebrospinal fluid abdominal cyst. Lee TG, et al. **Radiology** 127(1):220, Apr 78

Acute abdominal pain precipitated by phenformin–associated lactic acidosis. Bartecchi CE, et al. **South Med J** 71(3):344–6, Mar 78

The value of duodenal bile examination in the evaluation of persistent pain in the upper part of the abdomen. Humphries TJ, et al. **Surg Gynecol Obstet** 147(2):177–81, Aug 78

### ANATOMY & HISTOLOGY

Fetal weight estimation by ultrasonic measurement of abdominal circumference. Kearney K, et al. **Obstet Gynecol** 51(2):156–62, Feb 78

Reexamination of the deleterious effects of gastrointestinal contrast material on abdominal echography. Sarti DA, et al. **Radiology** 126(1):231–2, Jan 78

### BLOOD SUPPLY

Effect of morphine on splanchnic blood flow. Leaman DM, et al. **Br Heart J** 40(5):569–71, May 78

The use of prostaglandin E1 for enhanced visualization of the splanchnic circulation. Jonsson K, et al. **Radiology** 125(2):373–8, Nov 77

Source: Volume 9, 1978 *Cumulated Abridged Index Medicus*, photo taken at UK Libraries, Date: 2/20/2020

# MeSH Hierarchy: Example

1. Anatomy [A] (MeSH Branch)
    1.1. Body Regions [A01] <- Tree number
        1.1.1. Torso [A01.923]
            1.1.1.1. Abdomen [A01.923.047]

Abdomen may take one of 15 qualifiers, currently, including the qualifier: **/blood supply**.

The forward slash indicates that this is a qualifier.

# Nowadays

> Br Heart J. 1978 May;40(5):569-71. doi: 10.1136/hrt.40.5.569.

## Effect of morphine on splanchnic blood flow

D M Leaman, L Levenson, R Zelis, R Shiroff

PMID: 656226 PMCID: PMC483447 DOI: 10.1136/hrt.40.5.569

Free PMC article

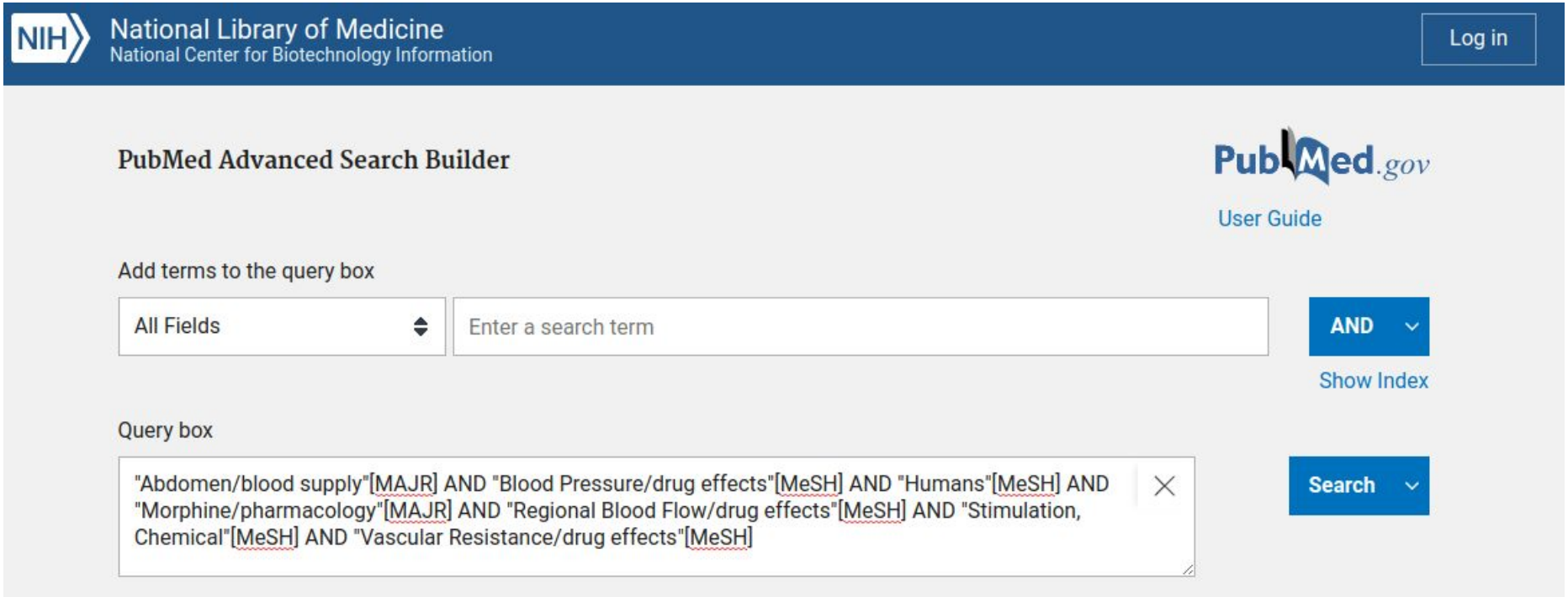
### MeSH terms

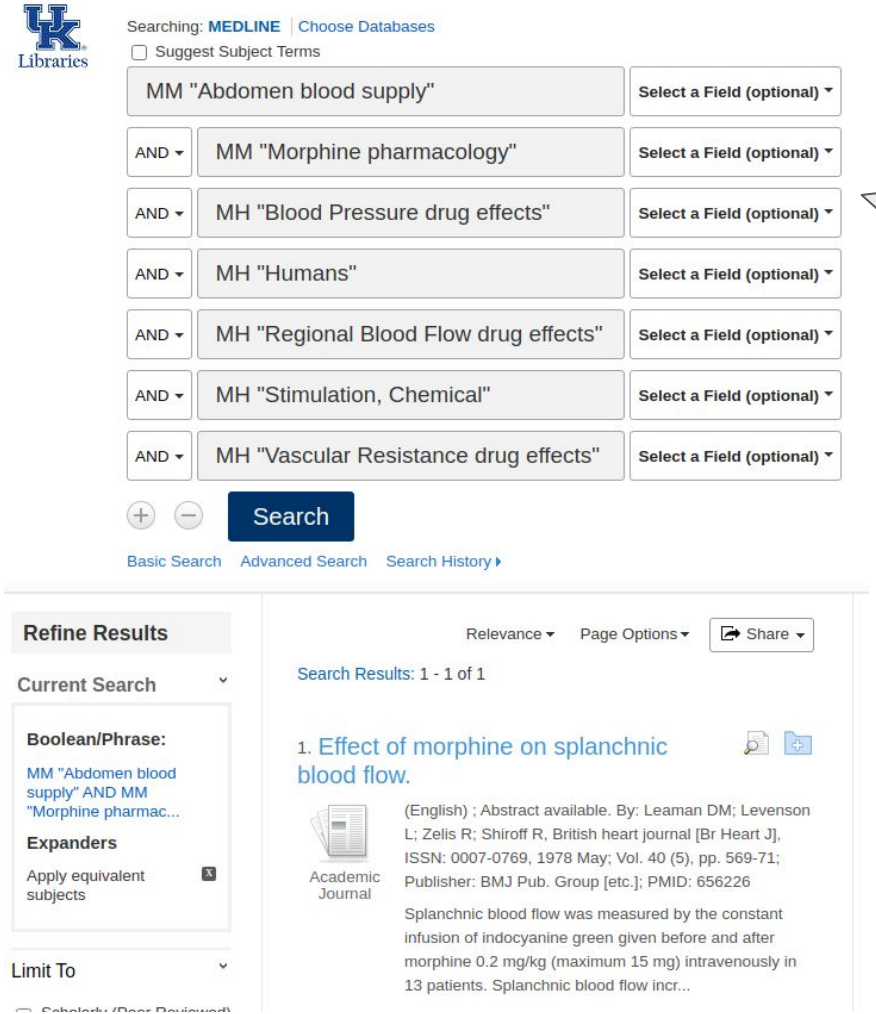
- Abdomen / blood supply*
- Blood Pressure / drug effects
- Humans
- Morphine / pharmacology*
- Regional Blood Flow / drug effects
- Stimulation, Chemical
- Vascular Resistance / drug effects

The asterisk indicates this is a major topic of the work.

As opposed to these other indexed headings /subheadings (aka, descriptors /qualifiers)

# To search PubMed based on previous example






Different MEDLINE platforms require different query syntaxes and apply different field codes. EBSCOhost does not require syntactical differences, like the forward slash, when including MeSH descriptors and qualifiers in the search field:

MM = Exact Major Subject Heading
MH = Exact Subject Heading

Whereas PubMed uses [MAJR] and [MeSH] along with the forward slash to represent qualifiers.

"abdomen/blood supply"[MAJR]
"blood pressure/drug effects"[MeSH]

# Converting queries to a single string

**In PubMed/MEDLINE:**

"Abdomen/blood supply"[MAJR] AND "Blood Pressure/drug effects"[MeSH] AND "Humans"[MeSH] AND "Morphine/pharmacology"[MAJR] AND "Regional Blood Flow/drug effects"[MeSH] AND "Stimulation, Chemical"[MeSH] AND "Vascular Resistance/drug effects"[MeSH]

**In EBSCOhost/MEDLINE:**

MM "Abdomen blood supply" AND MM "Morphine pharmacology" AND MH "Blood Pressure drug effects" AND MH "Humans" AND MH "Regional Blood Flow drug effects" AND MH "Stimulation, Chemical" AND MH "Vascular Resistance drug effects"

# MeSH Descriptors and Multiple Branches: *Dementia*

1. Diseases [C]
    1.1. Nervous System Diseases [C10]
        1.1.1. Central NSD [C10.228]
            1.1.1.1. Brain Diseases [C10.228.140]
                1.1.1.1.1. **Dementia** [C10.228.140.380]

1. Psychiatry and Psychology [F]
    1.1. Mental Disorders [F03]
        1.1.1. Neurocognitive Disorders [F03.615]
            1.1.1.1. **Dementia** [F03.615.400]

# MEDLINE Platform Study

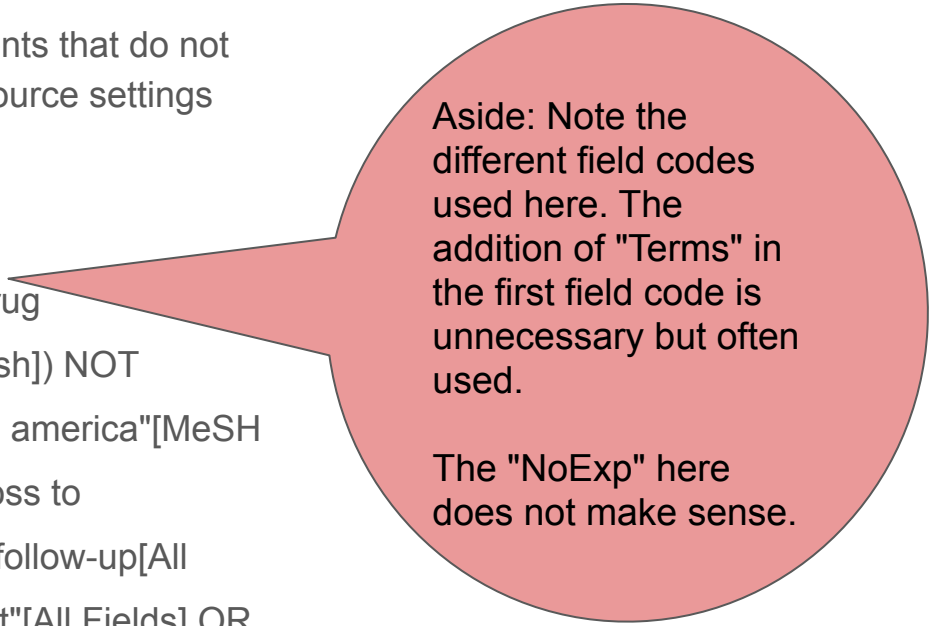
- Longitudinal study
  - 29 Query Cases
  - 5 Queries per Case
    - PubMed/MEDLINE
    - ProQuest/MEDLINE
    - EBSCOhost/MEDLINE
    - Web of Science/MEDLINE
    - Ovid/MEDLINE
  - Data collected once per month from Oct 2018 thru Sep 2019 (not including prior pilot data)
  - 145 Searches per month, and 1740 total searches for the study
  - Queries required syntactic changes but were designed to be:
    - Basic and straightforward
    - Semantically and logically equivalent across platforms
  - Queries were not designed to mimic real user needs or relevance

# Examples of real world queries not mimicked in this study:

# PubMed Search Strategy, Example 1

To locate literature investigating mortality of patients that do not follow up with antiretroviral treatments in low resource settings (https://doi.org/10.1371/journal.pone.0005790):

("2000/01/01"[PDAT]: "2008/12/31"[PDAT]) AND ("humans"[MeSH Terms] AND ("HIV Infections/drug therapy"[Mesh] OR "HIV Infections/mortality"[Mesh]) NOT ("Europe"[Mesh] OR "Australia"[Mesh] OR "north america"[MeSH Terms]) AND ("lost to follow-up"[All Fields] OR "loss to follow-up"[All Fields] OR (losses[All Fields] AND follow-up[All Fields]) OR "late patients"[All Fields] OR "dropout"[All Fields] OR "drop-out"[All Fields])) NOT "Clinical Trial "[Publication Type:NoExp]



# PubMed Search Strategy, Example 2

To locate all records that are systematic reviews (https://www.nlm.nih.gov/bsd/pubmed_subsets/sysreviews_strategy.html):

(((systematic review[ti] OR systematic literature review[ti] OR systematic scoping review[ti] OR systematic narrative review[ti] OR systematic qualitative review[ti] OR systematic evidence review[ti] OR systematic quantitative review [ti] OR systematic meta-review[ti] OR systematic critical review[ti] OR systematic mixed studies review[ti] OR systematic mapping review[ti] OR systematic cochrane review[ti] OR systematic search and review[ti] OR systematic integrative review[ti]) NOT comment[pt] NOT (protocol[ti] OR protocols[ti])) NOT MEDLINE [sb]) OR systematic review[pt]

Note that this search is nested up to three layers.

Aside: This search strategy includes a search for 'systematic reviews' as publication types. In MEDLINE, this "should" be sufficient to retrieve only systematic reviews. But this is a search strategy for all of PubMed (see NOT MEDLINE [sb]), and publication types are assigned only after items are indexed in MEDLINE.

# Examples from our study: PubMed/MEDLINE

| | |
|---|---|
| `"neoplasms"[ALL] AND 1950:2015[DP] AND medline[SB]` | *Most tag field* keyword search restricted by publication date |
| `"dementia"[MESH TERMS:NOEXP] AND 1950:2015[DP]` | MeSH term search, non-exploding, restricted by publication date |
| `"neoplasms"[MH] AND "immune"[ALL] AND 1950:2015[DP]` | MeSH term, exploding, and *most tag field* keyword search restricted by publication date |

# Exploding MeSH Descriptors

1. Diseases [C]
    1.1. Nervous System Diseases [C10]
        1.1.1. Central NSD [C10.228]
            1.1.1.1. Brain Diseases [C10.228.140]
                1.1.1.1.1. Dementia [C10.228.140.380]

"Brain Diseases"[MeSH] =
1,219,986 results as of 9/8/2020

"Brain Diseases"[MeSH:NoExp]
= 54,302 results as of 9/8/2020

If I want to search using this MeSH descriptor and also search all the more specific descriptors, then:

"Brain Diseases"[MeSH]

If I only want records returned with this MeSH descriptor and that do not include the more specific descriptors:

"Brain Diseases"[MeSH:NoExp]

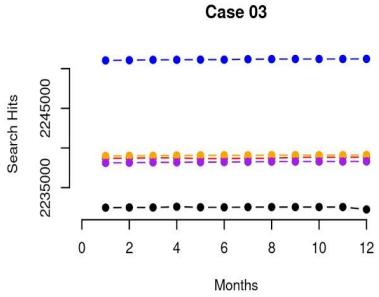
## Case #3: ALL FIELD KEYWORD and DATE 1950-2015

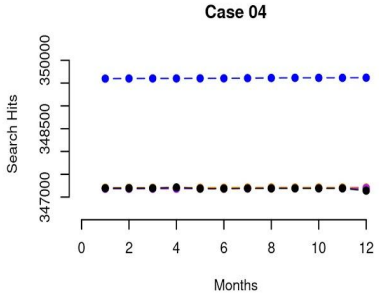
| Platform | Search Strategy | Hits, 10/19 |
|---|---|---|
| PubMed/MEDLINE | "neoplasms"[ALL] AND 1950:2015[DP] AND Medline[SB] | 2251033 |
| ProQuest/MEDLINE | NOFT("neoplasms") AND YR(1950-2015) | 2238698 |
| EBSCOhost/MEDLINE | TX("neoplasms") AND YR 1950-2015 | 2238997 |
| Web of Science/MEDLINE | TS=("neoplasms") AND PY=(1950-2015) | 2238118 |
| Ovid/MEDLINE | 1. neoplasms.AF 2. limit 1 to YR=1950-2015 | 2232480 |

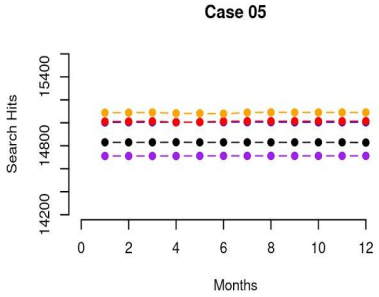
## Case #9: SINGLE MESH TERM, SINGLE BRANCH, EXPLODE, PLUS ALL FIELD KEYWORD, PUB DATE 1950-2015

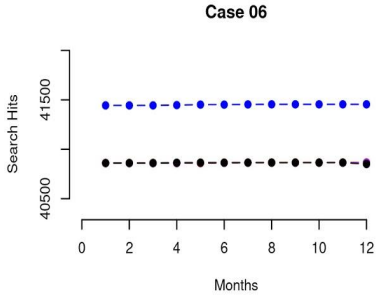
| Platform | Search Strategy | Hits, 10/19 |
|---|---|---|
| PubMed/MEDLINE | "neoplasms"[MH] AND "immune"[ALL] AND 1950:2015[DP] | 72297 |
| ProQuest/MEDLINE | MESH.EXPLODE("neoplasms") AND NOFT("immune") AND YR(1950-2015) | 72641 |
| EBSCOhost/MEDLINE | MH("neoplasms+") AND TX("immune") AND YR 1950-2015 | 72987 |
| Web of Science/MEDLINE | MH:exp=("neoplasms") AND TS=("immune") AND PY=(1950-2015) | 14711 |
| Ovid/MEDLINE | 1. EXP neoplasms/ AND immune.AF 2. limit 1 to YR=1950-2015 | 71594 |

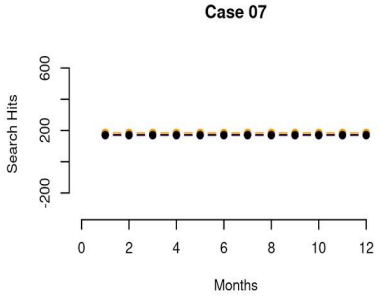
# Macro View: Queries limited by Publication Date

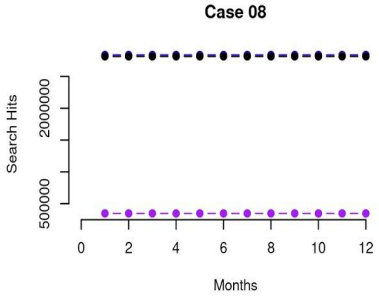

Case 03
Search Hits
2235000
2245000
0
2
4
6
8
10
12
Months

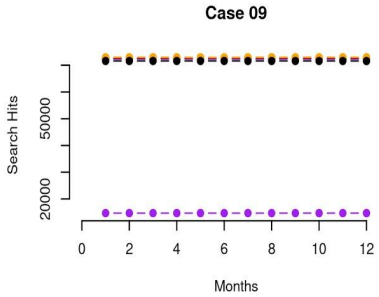

Case 04
Search Hits
347000
348500
350000
0
2
4
6
8
10
12
Months

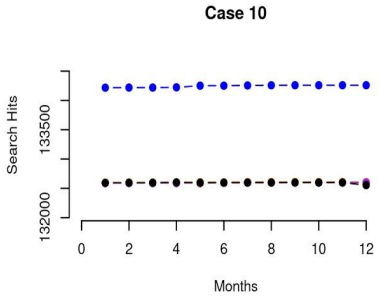

Case 05
Search Hits
14200
14800
15400
0
2
4
6
8
10
12
Months

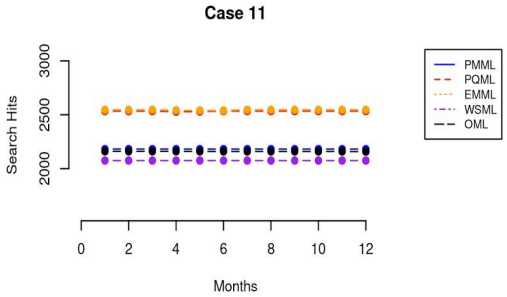

Case 06
Search Hits
40500
41500
0
2
4
6
8
10
12
Months

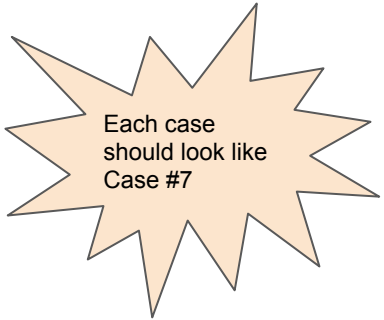

Case 07
Search Hits
-200
200
600
0
2
4
6
8
10
12
Months


Case 08
Search Hits
500000
2000000
0
2
4
6
8
10
12
Months

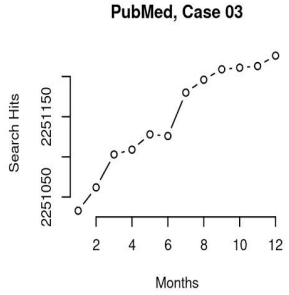

Case 09
Search Hits
20000
50000
0
2
4
6
8
10
12
Months

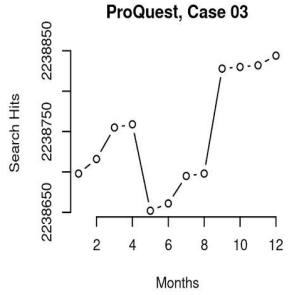

Case 10
Search Hits
132000
133500
0
2
4
6
8
10
12
Months

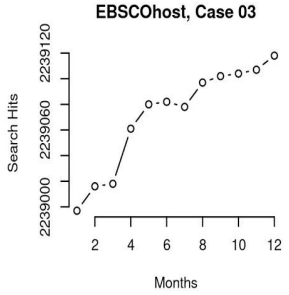

Case 11
Search Hits
2000
2500
3000
0
2
4
6
8
10
12
Months
PMML
PQML
EMML
WSML
OML

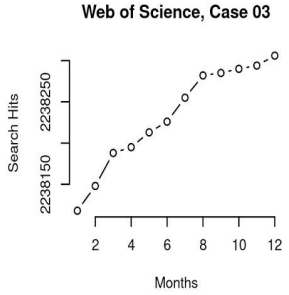

Each case
should look like
Case #7

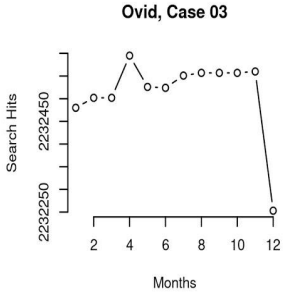
# Close Up View: Queries limited by Publication Date

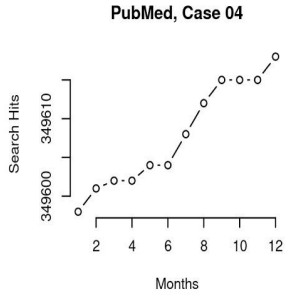

PubMed, Case 03
Search Hits
2251050
2251150
2
4
6
8
10
12
Months

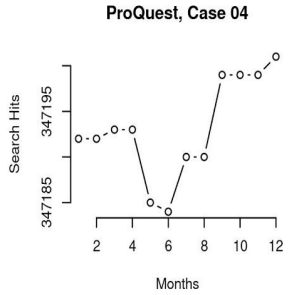

ProQuest, Case 03
Search Hits
2238650
2238750
2238850
2
4
6
8
10
12
Months

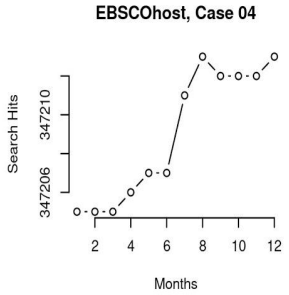

EBSCOhost, Case 03
Search Hits
2239000
2239060
2239120
2
4
6
8
10
12
Months

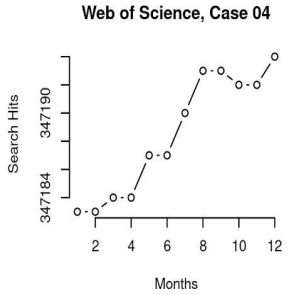

Web of Science, Case 03
Search Hits
2238150
2238250
2
4
6
8
10
12
Months

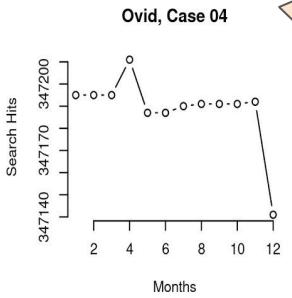

Ovid, Case 03
Search Hits
2232250
2232450
2
4
6
8
10
12
Months

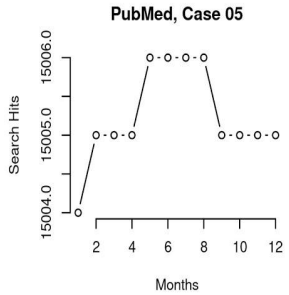

PubMed, Case 04
Search Hits
349600
349610
2
4
6
8
10
12
Months

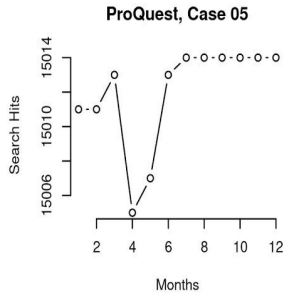

ProQuest, Case 04
Search Hits
347185
347195
2
4
6
8
10
12
Months

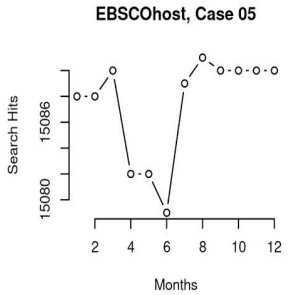

EBSCOhost, Case 04
Search Hits
347206
347210
2
4
6
8
10
12
Months

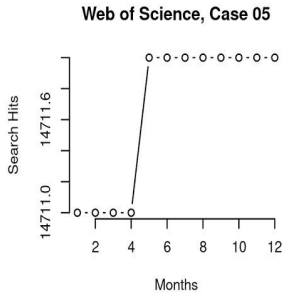

Web of Science, Case 04
Search Hits
347184
347190
2
4
6
8
10
12
Months

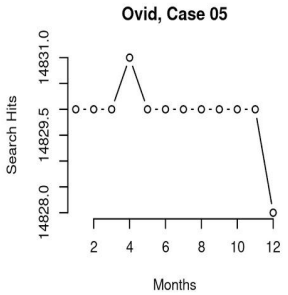

Ovid, Case 04
Search Hits
347140
347170
347200
2
4
6
8
10
12
Months

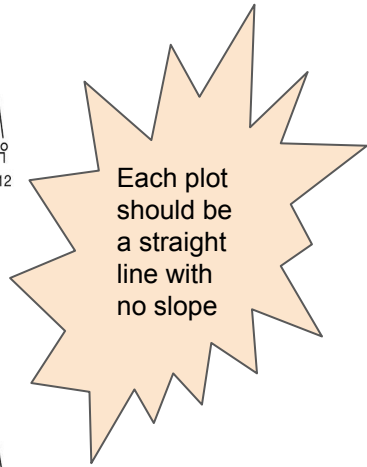

PubMed, Case 05
Search Hits
15004.0
15005.0
15006.0
2
4
6
8
10
12
Months

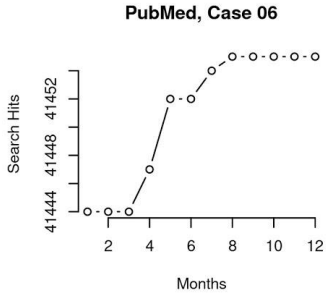

ProQuest, Case 05
Search Hits
15006
15010
15014
2
4
6
8
10
12
Months

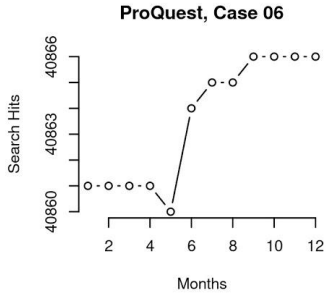

EBSCOhost, Case 05
Search Hits
15080
15086
2
4
6
8
10
12
Months

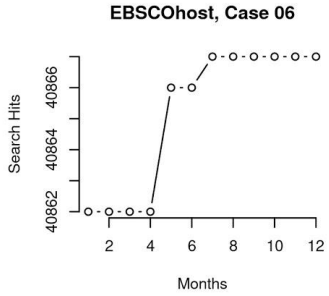

Web of Science, Case 05
Search Hits
14711.0
14711.6
2
4
6
8
10
12
Months

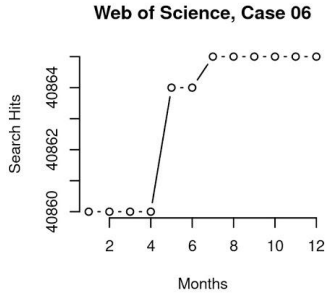

Ovid, Case 05
Search Hits
14828.0
14829.5
14831.0
2
4
6
8
10
12
Months

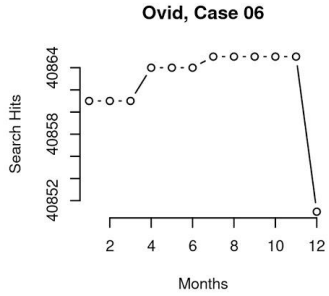

Each plot
should be
a straight
line with
no slope

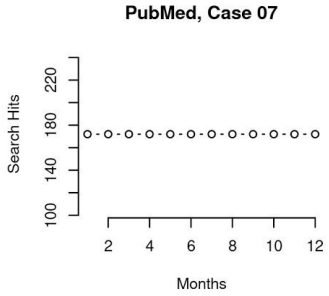

PubMed, Case 06
Search Hits
41444
41448
41452
2
4
6
8
10
12
Months

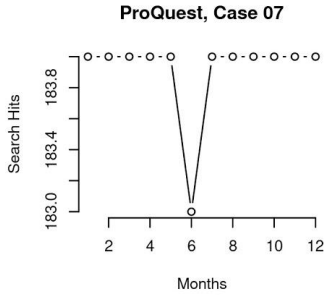

ProQuest, Case 06
Search Hits
40860
40863
40866
2
4
6
8
10
12
Months

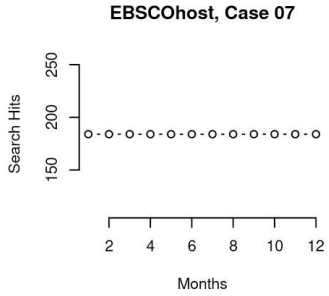

EBSCOhost, Case 06
Search Hits
40862
40864
40866
2
4
6
8
10
12
Months

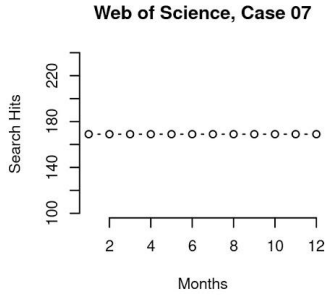

Web of Science, Case 06
Search Hits
40860
40862
40864
2
4
6
8
10
12
Months

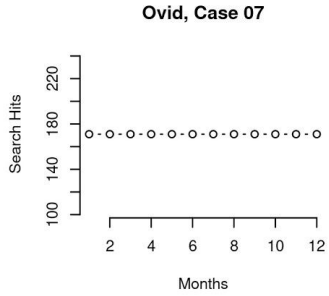

Ovid, Case 06
Search Hits
40852
40858
40864
2
4
6
8
10
12
Months

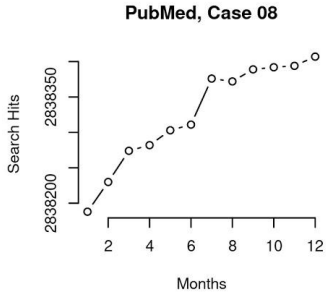

PubMed, Case 07
Search Hits
100
140
180
220
2
4
6
8
10
12
Months

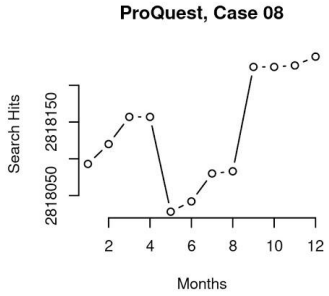

ProQuest, Case 07
Search Hits
183.0
183.4
183.8
2
4
6
8
10
12
Months

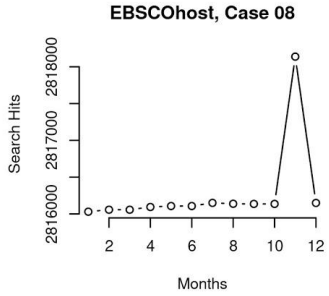

EBSCOhost, Case 07
Search Hits
150
200
250
2
4
6
8
10
12
Months

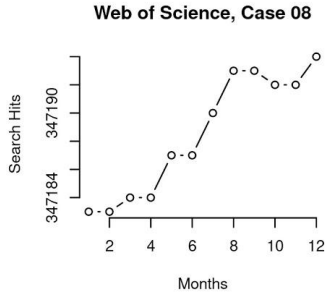

Web of Science, Case 07
Search Hits
100
140
180
220
2
4
6
8
10
12
Months

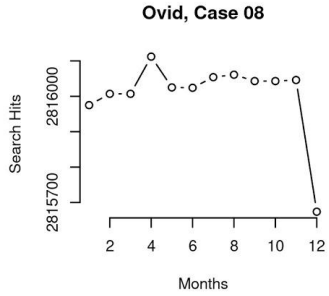

Ovid, Case 07
Search Hits
100
140
180
220
2
4
6
8
10
12
Months

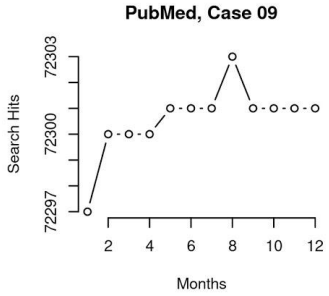

PubMed, Case 08
Search Hits
2838200
2838350
2
4
6
8
10
12
Months

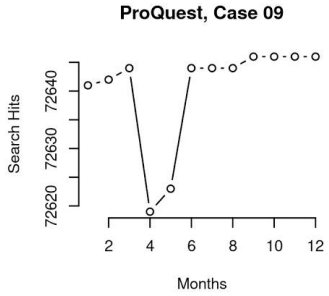

ProQuest, Case 08
Search Hits
2818050
2818150
2
4
6
8
10
12
Months

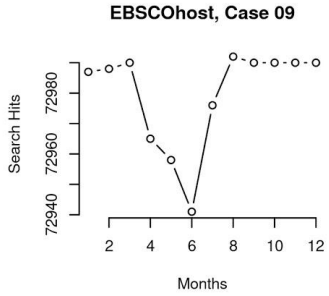

EBSCOhost, Case 08
Search Hits
2816000
2817000
2818000
2
4
6
8
10
12
Months

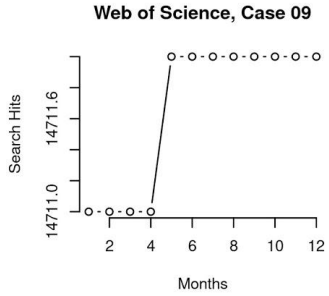

Web of Science, Case 08
Search Hits
347184
347190
2
4
6
8
10
12
Months

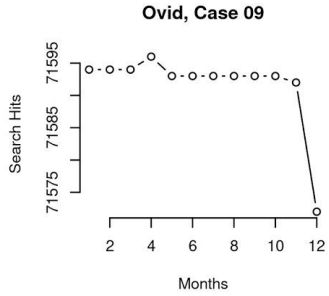

Ovid, Case 08
Search Hits
2815700
2816000
2
4
6
8
10
12
Months

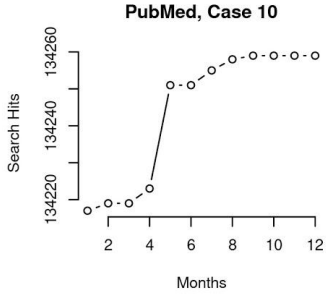

PubMed, Case 09
Search Hits
72297
72300
72303
2
4
6
8
10
12
Months

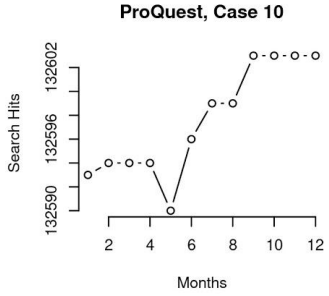

ProQuest, Case 09
Search Hits
72620
72630
72640
2
4
6
8
10
12
Months

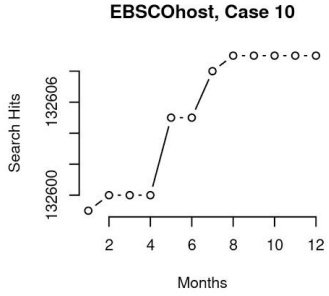

EBSCOhost, Case 09
Search Hits
72940
72960
72980
2
4
6
8
10
12
Months

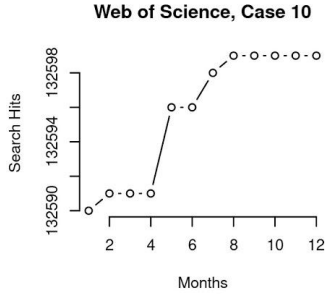

Web of Science, Case 09
Search Hits
14711.0
14711.6
2
4
6
8
10
12
Months

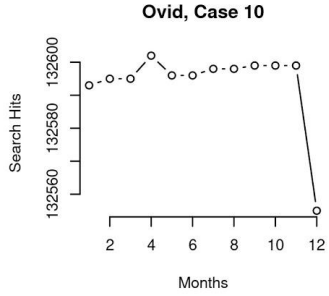

Ovid, Case 09
Search Hits
71575
71585
71595
2
4
6
8
10
12
Months

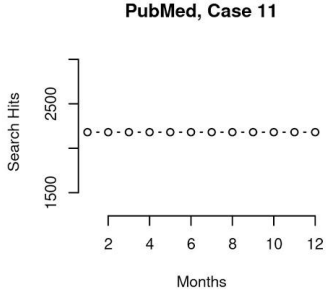

PubMed, Case 10
Search Hits
134220
134240
134260
2
4
6
8
10
12
Months

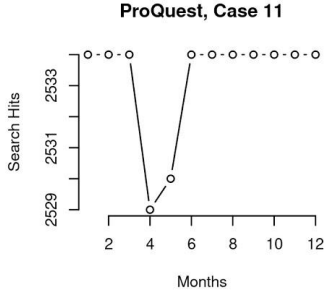

ProQuest, Case 10
Search Hits
132590
132596
132602
2
4
6
8
10
12
Months

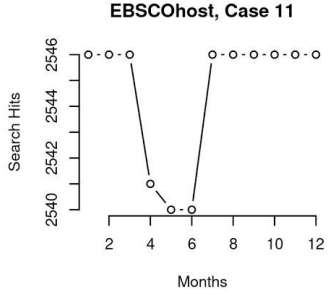

EBSCOhost, Case 10
Search Hits
132600
132606
2
4
6
8
10
12
Months


Web of Science, Case 10
Search Hits
132590
132594
132598
2
4
6
8
10
12
Months

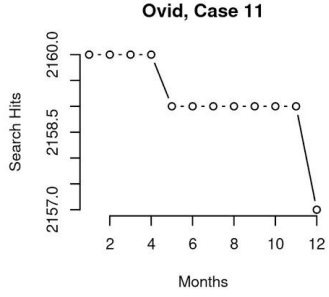

Ovid, Case 10
Search Hits
132560
132580
132600
2
4
6
8
10
12
Months


PubMed, Case 11
Search Hits
1500
2500
2
4
6
8
10
12
Months

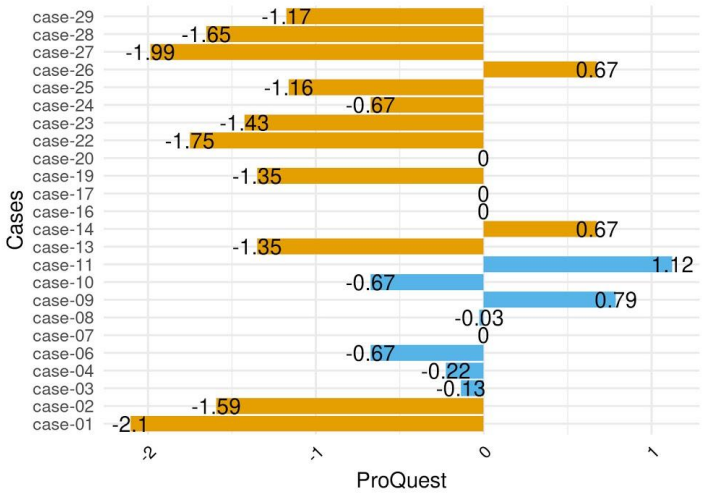

ProQuest, Case 11
Search Hits
2529
2531
2533
2
4
6
8
10
12
Months

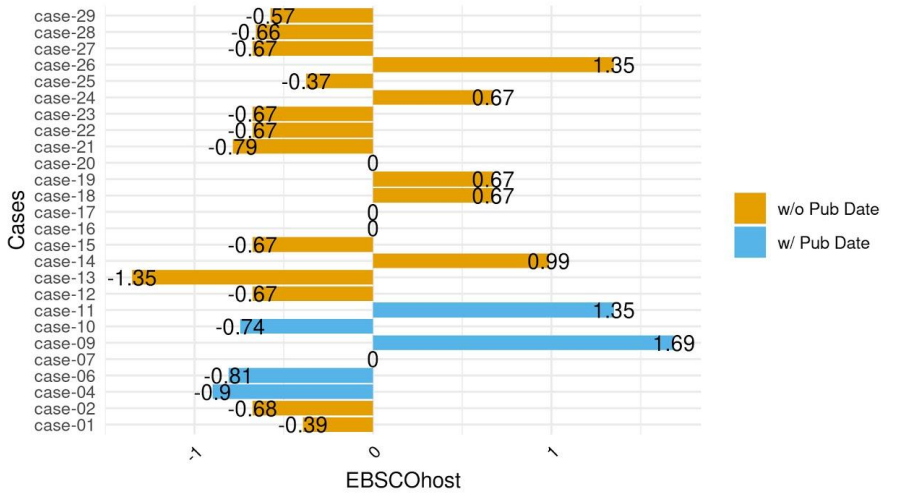

EBSCOhost, Case 11
Search Hits
2540
2542
2544
2546
2
4
6
8
10
12
Months

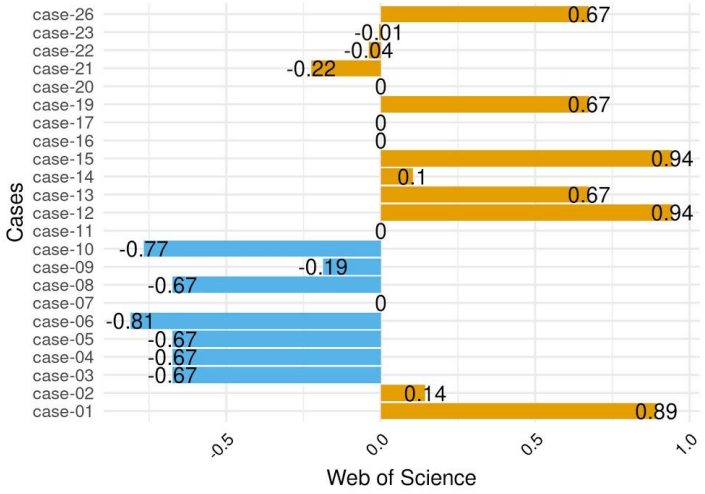
Web of Science, Case 11
Search Hits
1500
2000
2500
2
4
6
8
10
12
Months

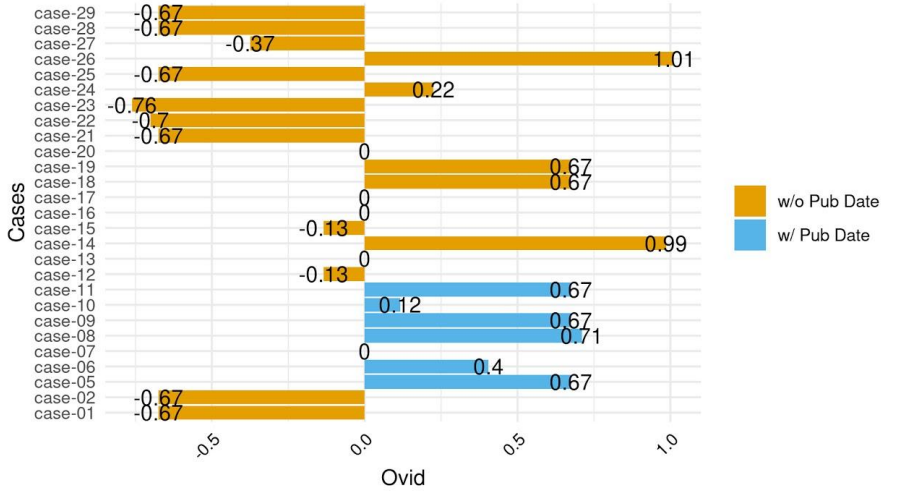

Ovid, Case 11
Search Hits
2157.0
2158.5
2160.0
2
4
6
8
10
12
Months

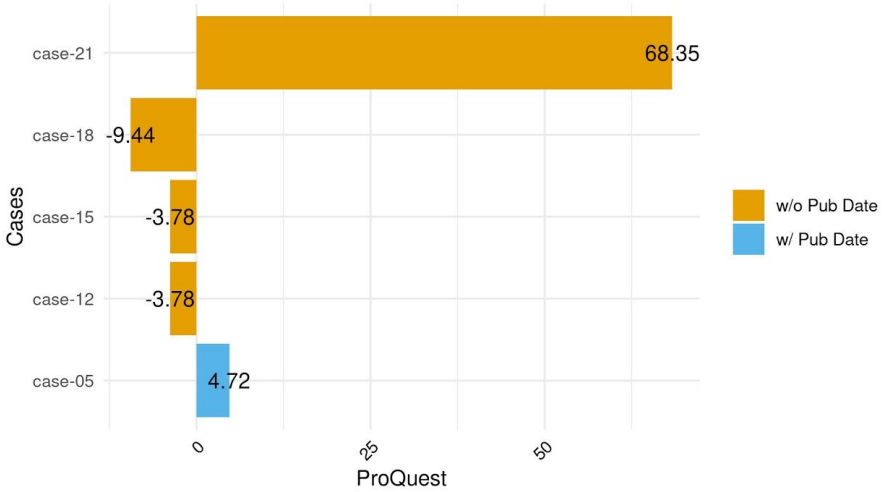
# Comparison to PubMed/MEDLINE

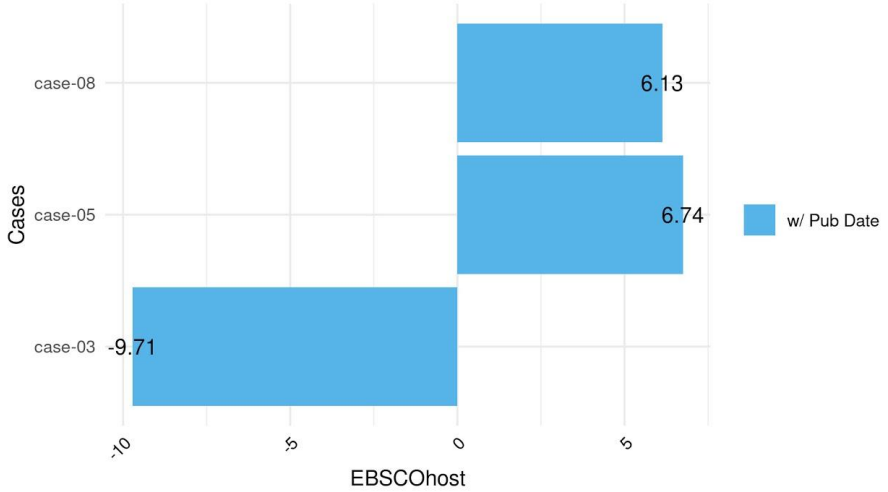

Cases
case-29 -1.17
case-28 -1.65
case-27 -1.99
case-26 0.67
case-25 -1.16
case-24 -0.67
case-23 -1.43
case-22 -1.75
case-20 0
case-19 -1.35
case-17 0
case-16 0
case-14 0.67
case-13 -1.35
case-11 1.12
case-10 -0.67
case-09 0.79
case-08 -0.03
case-07 0
case-06 -0.67
case-04 -0.22
case-03 -0.13
case-02 -1.59
case-01 -2.1
-2
-1
0
1
ProQuest
w/o Pub Date
w/ Pub Date

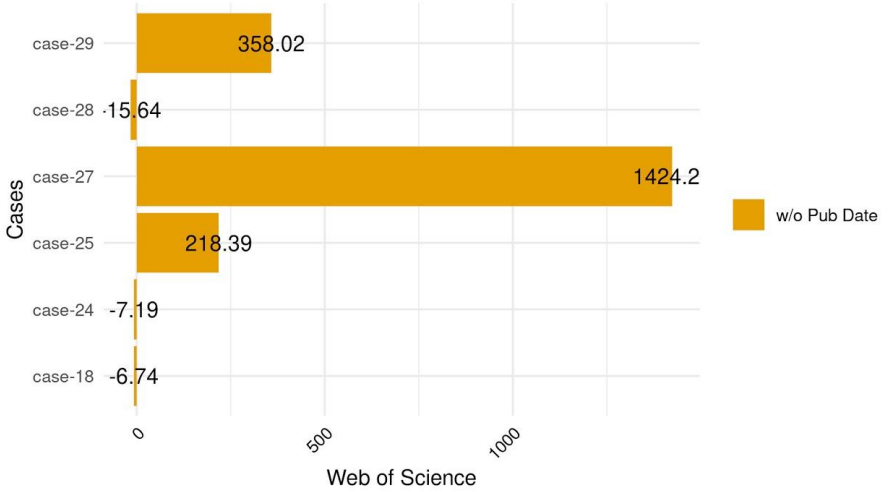

Cases
case-29 -0.57
case-28 -0.66
case-27 -0.67
case-26 1.35
case-25 -0.37
case-24 0.67
case-23 -0.67
case-22 -0.67
case-21 -0.79
case-20 0
case-19 0.67
case-18 0.67
case-17 0
case-16 0
case-15 -0.67
case-14 0.99
case-13 -1.35
case-12 -0.67
case-11 1.35
case-10 -0.74
case-09 1.69
case-07 0
case-06 -0.81
case-04 -0.9
case-02 -0.68
case-01 -0.39
-1
0
1
EBSCOhost
w/o Pub Date
w/ Pub Date

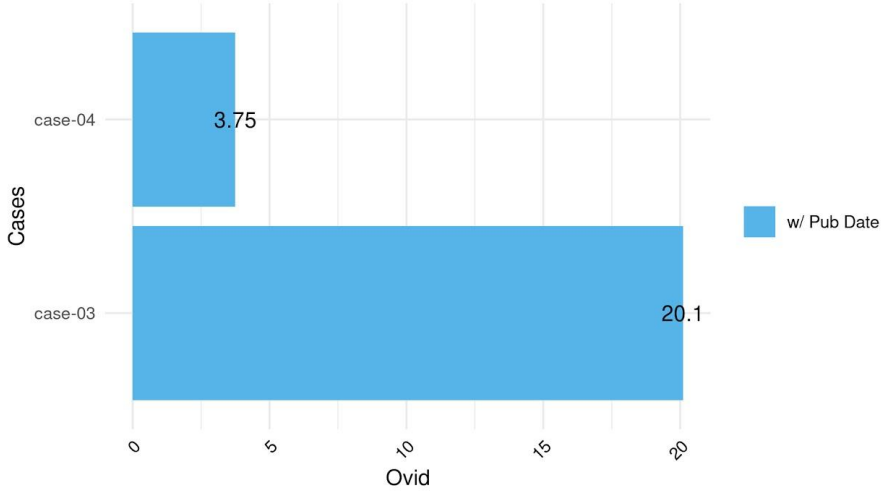

Cases
case-26 0.67
case-23 -0.01
case-22 -0.04
case-21 -0.22
case-20 0
case-19 0.67
case-17 0
case-16 0
case-15 0.94
case-14 0.1
case-13 0.67
case-12 0.94
case-11 0
case-10 -0.77
case-09 -0.19
case-08 -0.67
case-07 0
case-06 -0.81
case-05 -0.67
case-04 -0.67
case-03 -0.67
case-02 0.14
case-01 0.89
-0.5
0.0
0.5
1.0
Web of Science
w/o Pub Date
w/ Pub Date


Cases
case-29 -0.67
case-28 -0.67
case-27 -0.37
case-26 1.01
case-25 -0.67
case-24 0.22
case-23 -0.76
case-22 -0.7
case-21 -0.67
case-20 0
case-19 0.67
case-18 0.67
case-17 0
case-16 0
case-15 -0.13
case-14 0.99
case-13 0
case-12 -0.13
case-11 0.67
case-10 0.12
case-09 0.67
case-08 0.71
case-07 0
case-06 0.4
case-05 0.67
case-02 -0.67
case-01 -0.67
-0.5
0.0
0.5
1.0
Ovid
w/o Pub Date
w/ Pub Date


case-21
case-18
case-15
case-12
case-05
68.35
-9.44
-3.78
-3.78
4.72
Cases
0
25
50
ProQuest
w/o Pub Date
w/ Pub Date


case-08
case-05
case-03
6.13
6.74
-9.71
Cases
-10
-5
0
5
EBSCOhost
w/ Pub Date


case-29
case-28
case-27
case-25
case-24
case-18
358.02
-15.64
1424.2
218.39
-7.19
-6.74
Cases
0
500
1000
Web of Science
w/o Pub Date

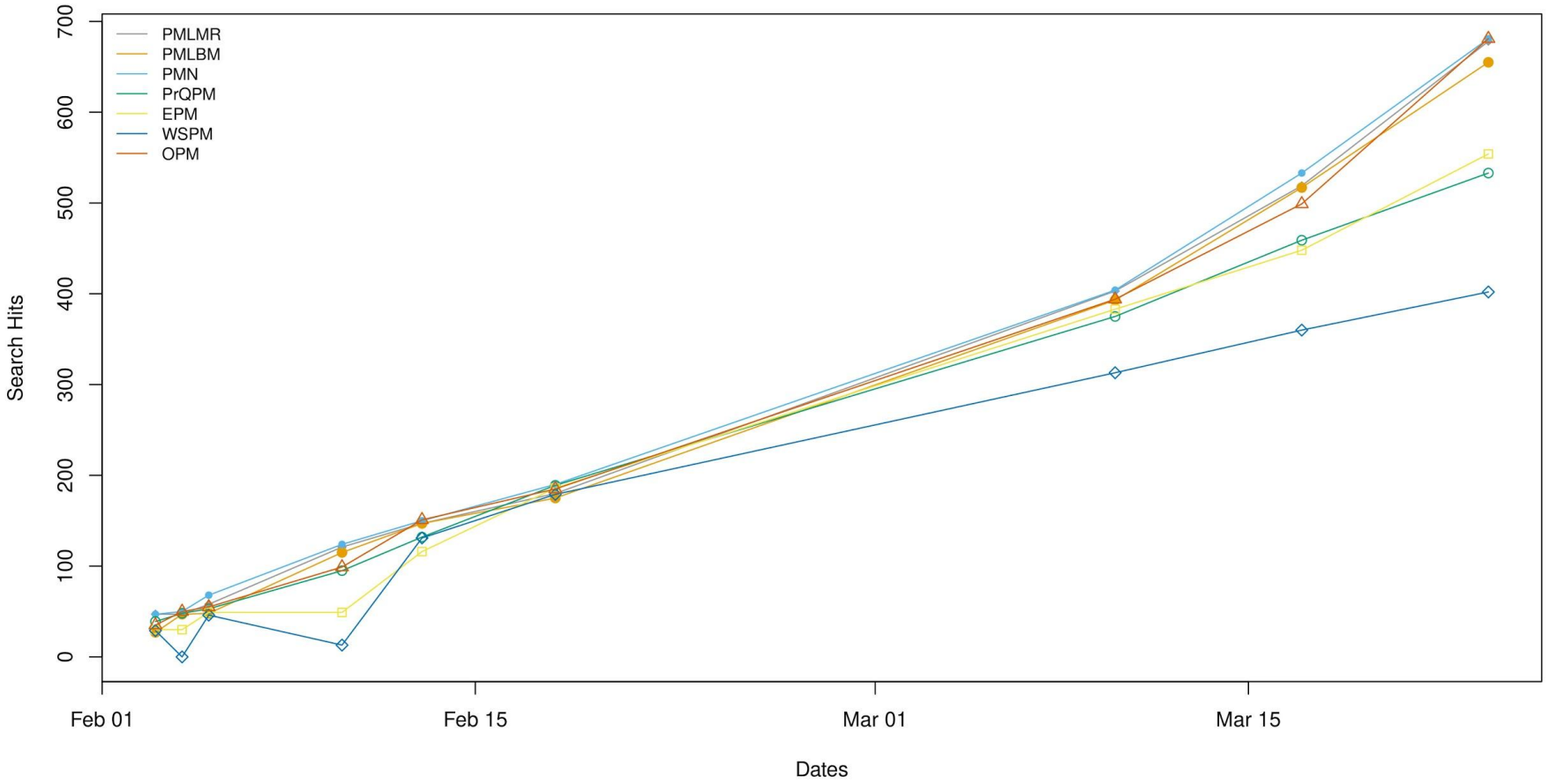

case-04
case-03
3.75
20.1
Cases
0
5
10
15
20
Ovid
w/ Pub Date

# Covid-19 & PubMed

- The NLM issued a preliminary search strategy on January 24, 2020 for new research on COVID-19:

```
2019-nCoV[All Fields] OR (wuhan[tiab] AND coronavirus[tiab])
```

- We monitored seven versions of PubMed to compare time delays among the platforms
- The overall gist is PubMed is the platform of choice if timeliness is an issue due to substantial delays in data migration from PubMed to other PubMed/MEDLINE information service providers even if PubMed is not generally the best platform of choice [ we don't know which is best yet ]


PMLMR
PMLBM
PMN
PrQPM
EPM
WSPM
OPM
Search Hits
0
100
200
300
400
500
600
700
Feb 01
Feb 15
Mar 01
Mar 15
Dates

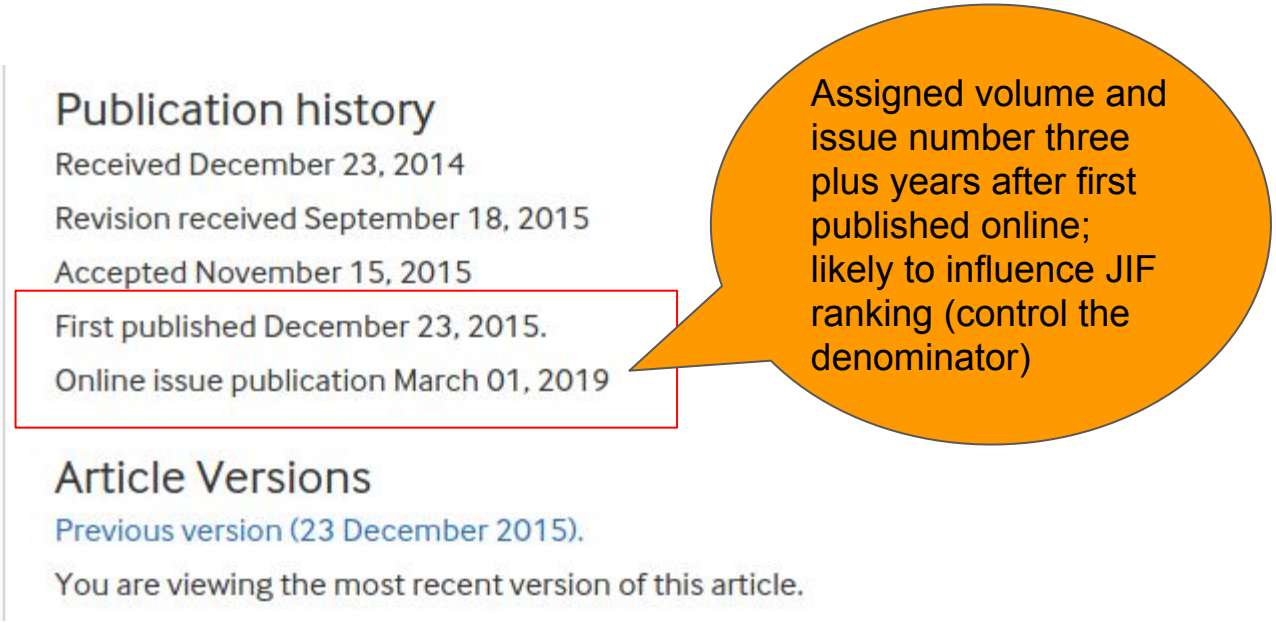
# Problems

- Web of Science/MEDLINE is broken
- Database currency due to lag in updates across platforms
- Exploding is inconsistent
- Online first and print publications reduce effectiveness of *bibliographic control*, a problem caused by the migration from print to digital scholarly communication and publishing

# Growth after Publication Date

Source: *BMJ Supportive & Palliative Care*
https://spcare.bmj.com/content/early/2015/12/23/bmjspcare-2014-000835.info?versioned=true

# Growth after Publication Date




Source: *BMJ Supportive & Palliative Care*
https://spcare.bmj.com/content/9/1/67.info

# Future Research

- Some complications appear to due to systems combining Boolean models and vector models of information retrieval.
  - Boolean model is based on set theory
    - This should be relatively straightforward. Records contain or do not contain MeSH terms. But this is complicated by the use of MeSH as a thesaurus and the ability to explode terms and include narrower subsets. Also, it's not clear if the platforms are structuring or indexing the bibliographic records correctly (data integrity) or if the platforms are exploding descriptors properly.
  - Vector model is based on weighting terms, such as by word (term) frequency and with respect to inverse document frequency: a term is more important the more frequently it appears in a document and then less frequently it appears in the document collection. PubMed has adapted this and applied various machine learning algorithms based on signal detection in their new version of Best Match. It's not clear how other platforms apply vector or like models, since they are proprietary.

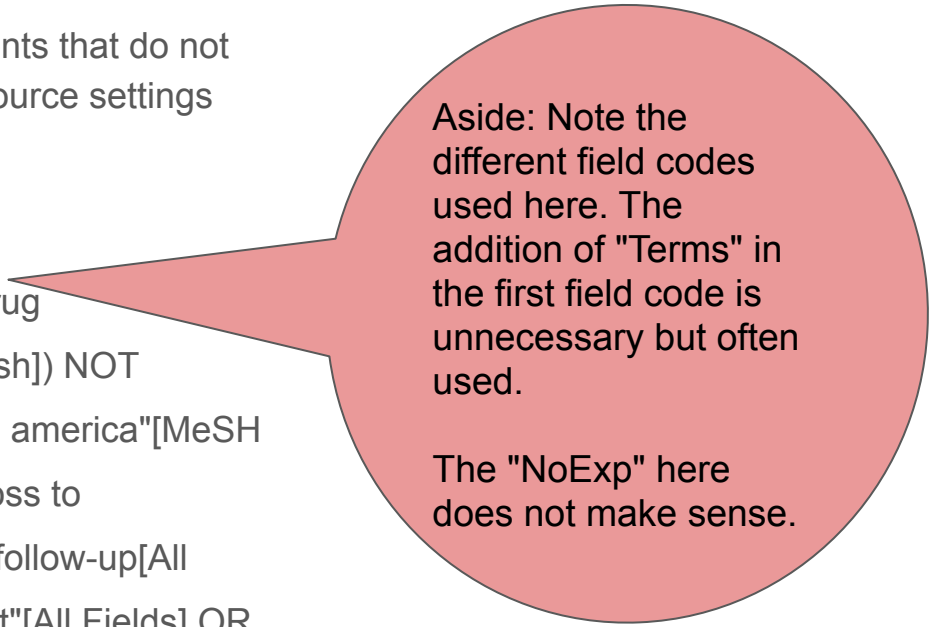
# Future Research

- None of our peer reviews have commented on our search strategies
- Search strategies are often confusing (see example real work queries)
- Need research on disentangling these
  - Might draw from programming language style guides:
    - PEP 8 -- Style Guide for Python Code: https://www.python.org/dev/peps/pep-0008/
    - Advanced R: https://adv-r.hadley.nz/index.html
    - Google Style Guides: https://google.github.io/styleguide/
    - Work on developing lint software to check for bugs, style issues, and other error types in search strategies: https://en.wikipedia.org/wiki/Lint_(software)

# PubMed Search Strategy, Example 1

To locate literature investigating mortality of patients that do not follow up with antiretroviral treatments in low resource settings (https://doi.org/10.1371/journal.pone.0005790):

("2000/01/01"[PDAT]: "2008/12/31"[PDAT]) AND ("humans"[MeSH Terms] AND ("HIV Infections/drug therapy"[Mesh] OR "HIV Infections/mortality"[Mesh]) NOT ("Europe"[Mesh] OR "Australia"[Mesh] OR "north america"[MeSH Terms]) AND ("lost to follow-up"[All Fields] OR "loss to follow-up"[All Fields] OR (losses[All Fields] AND follow-up[All Fields]) OR "late patients"[All Fields] OR "dropout"[All Fields] OR "drop-out"[All Fields])) NOT "Clinical Trial "[Publication Type:NoExp]



# PubMed Search Strategy, Example 2

To locate all records that are systematic reviews (https://www.nlm.nih.gov/bsd/pubmed_subsets/sysreviews_strategy.html):

(((systematic review[ti] OR systematic literature review[ti] OR systematic scoping review[ti] OR systematic narrative review[ti] OR systematic qualitative review[ti] OR systematic evidence review[ti] OR systematic quantitative review [ti] OR systematic meta-review[ti] OR systematic critical review[ti] OR systematic mixed studies review[ti] OR systematic mapping review[ti] OR systematic cochrane review[ti] OR systematic search and review[ti] OR systematic integrative review[ti]) NOT comment[pt] NOT (protocol[ti] OR protocols[ti])) NOT MEDLINE [sb]) OR systematic review[pt]

Note that this search is nested up to three layers.

Aside: This search strategy includes a search for 'systematic reviews' as publication types. In MEDLINE, this "should" be sufficient to retrieve only systematic reviews. But this is a search strategy for all of PubMed (see NOT MEDLINE [sb]), and publication types are assigned only after items are indexed in MEDLINE.