Added value of MeSH terms in search strategies of systematic reviews

Victor Leblanc, Aghiles Hamroun, Raphaël Bentegeac, Bastien Le Guellec, Rémi Lenain, Emmanuel Chazard

Submitted to: Journal of Medical Internet Research
on: October 18, 2023

Table of Contents

Added value of MeSH terms in search strategies of systematic reviews

Victor Leblanc[1]; Aghiles Hamroun[1]; Raphaël Bentegeac[1]; Bastien Le Guellec[1]; Rémi Lenain[1]; Emmanuel Chazard[2]

[1]Public Health Department CHU Lille, Université de Lille, France Lille FR
[2]ULR 2694 Metrics, CERIM Public Health Department CHU Lille, Université de Lille Lille FR

Corresponding Author:
Victor Leblanc
Public Health Department
CHU Lille, Université de Lille, France
42 Rue Paul Duez
Lille
FR

Abstract

Background: The massive increase in the number of publications enhances scientific and medical knowledge but makes it more complicated to summarize research results. The MeSH thesaurus was created in the mid-20th century with the aim of systematizing indexing and facilitating publication retrieval. Despite the advent of search engines, few studies have questioned the relevance of the MeSH thesaurus and none has done so in a systematic manner.

Objective: The objective of the present work was to evaluate the utility of MeSH terms in Pubmed queries for systematic literature reviews.

Methods: We selected systematic literature reviews published in four prestigious journals between 2012 and 2021 and for which a PubMed query was provided. Each original query (V1) was transformed to obtain a version with free text terms only (V2) and a version with MeSH terms only (V3). The three queries were compared with regard to their sensitivity and positive predictive value (PPV).

Results: A total of 70 systematic literature reviews were included. Three V1 queries (4.3%) contained MeSH terms only, 8 (11.4%) contained free text terms only, and 59 (84.3%) contained both MeSH terms and free text terms. The transition from V1 to V2 had no effect on the number of relevant articles retrieved for 39 of the 70 reviews (56%). The deletion of MeSH terms decreased the median sensitivity (from 50.0% for V1 to 42.4% for V2) and increased the median PPV (from 1.3% for V1 to 1.6% for V2). Conversely, the deletion of free text terms decreased the median sensitivity (from 50% for V1 to 16.7% for V3) and decreased the median PPV (from 1.6% for V2 to 1.0% for V3). In other words, queries containing MeSH terms and free text terms provide an average of 4 additional relevant papers per SR but an additional 769 papers have to be screened.

Conclusions: For researchers building a PubMed query for a systematic literature review, MeSH terms do not appear to be essential. In fact, MeSH terms sometimes slightly increase the number of relevant articles identified but significantly increase the workload required for filtering. Hence, MeSH terms should never be used alone.

(JMIR Preprints 18/10/2023:53781)
DOI: https://doi.org/10.2196/preprints.53781

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Added value of MeSH terms in search strategies of systematic reviews

Leblanc V., Hamroun A., Bentegeac R., Le Guellec B., Lenain R., Chazard E.

Corresponding Author: Victor Leblanc
leblancvictor59@gmail.com, +33 6 37 00 49 71
Public Health and Epidemiology Service
University Hospital Center
2 Av. Oscar Lambret, Lille, France

Introduction

The number of articles published in scientific and medical journals has been increasing exponentially since the late 20th century. In 2021 alone, over 1,700,000 indexed, full-text articles were included in the PubMed database. In response to the massive production of scientific knowledge, the need for access to synthetic scientific data has been driven by the emergence of evidence-based medicine (EBM) [1] and the establishment of national regulatory bodies, medical associations and learned societies that provide guidelines on best practice.

In this context, systematic literature reviews (SRs, a type of analysis developed in the 1970s) are becoming more important. Given that quality of an SR depends largely on the research methodology, building search queries is a crucial part of the review process. The challenge of constructing a query for a systematic review lies in the absolute necessity of being as sensitive as possible, despite the fact that this query will return at most a few tens of thousands of articles among the hundreds of millions that make up the scientific literature. [2].

In the mid-20th century, researchers started to develop a common vocabulary that facilitated article indexing and retrieval and helped to avoid misunderstandings [3–5]. These efforts led to the creation of the Medical Subject Headings (MeSH) thesaurus in the 1960s by the US National Library of Medicine (NLM) [6]. PubMed (the NLM's search engine), which is one of the most widely used search engines [7], heavily relies on MeSH terms to assist users in their literature searches. The MeSH thesaurus is intended to facilitate literature searches by limiting

term permutations [8,9]. In other words, it assigns a unique term to a concept – regardless of the language used or the time period concerned.

Subsequent improvements in search engine performance have enabled researchers to query databases with simple free-text terms, rather than MeSH terms. Furthermore, the massive influx of publications and the emergence of many new scientific and medical topics have led to delays in MeSH indexing and difficulties in updating the thesaurus [10]. Finally, although frequently recommended [11–13], the value of using the MeSH thesaurus in queries for literature reviews has never been systematically assessed. The few studies to have tested the utility of MeSH terms in SRs have limitations, such as a small sample size or a lack of generalizability [14–20]. Lastly, some studies simply compared the numbers of results retrieved for a given query but did not evaluate the results' relevance [21].

To the best our knowledge, only one study has extensively explored the relevance of MeSH terms with regard to the results of SRs [22]. The study concluded that the use of queries based on free-text alone (i.e. free-text terms) appeared to decrease the retrieval of articles of interest, relative to queries based on both free-text terms and MeSH terms. However, this study included SRs from a single research center, which limited the generalizability of the findings. Moreover, the MEDLINE database was queried with the Ovid search engine, rather than PubMed. We therefore decided to evaluate this question in more detail. The objective of the present work was to estimate the added value of using MeSH terms in PubMed queries for SRs.

Methods

Paper selection

We first selected the top six journals in the "Medicine. General & Internal" Journal Citation Reports category, according to the impact factors computed by Clarivate [23,24]. Next, we selected all the PubMed-indexed SRs published in the six journals between 2012 and 2021 and for which the free full text was available on PubMed Central. The time period was chosen arbitrarily, with the objective

of obtaining at least 60 SRs. The following PubMed query was used: ‘("The New England Journal of Medicine"[Journal] OR "Lancet London England"[Journal] OR "JAMA"[Journal] OR "Nature Reviews Disease Primers"[Journal] OR "BMJ Clinical Research Ed"[Journal] OR "Annals of Internal Medicine"[Journal]) AND "loattrfree full text"[Filter] AND 2012/01/01:2021/12/31[Date - Publication] AND systematic review[Filter]’.

The exclusion criteria were as follows: articles other than an SR, the absence of a published search query, the use of queries in multiple parts that had to be assembled, the absence of a query specifically built for PubMed, a query that did not return any results, a query that returned more than 100,000 results and a query with only MeSH terms or without MeSH terms. The sorting was carried out by a single researcher (VL).

Analysis of the PubMed results

The query was extracted from each included SR and inserted into the PubMed search bar. PubMed has a feature called automatic term mapping (ATM) [25]: when terms not enclosed in quotation marks are inserted in the search bar, they are automatically transformed into a query segment that contains several descriptors, such as [MeSH terms], [tiab], and [all fields]. To ensure greater reproducibility, we checked for the automatic transformation of queries. This step was important because PubMed’s ATM feature might add MeSH terms to query initially considered to be free of such terms. Hence, we always retrieved the query formatted by PubMed's ATM (henceforth referred to as V1; see Figure 1).

For each SR, V1 was transformed into a V2 query by replacing each MeSH term in the query with a free-text term that had to be present in the title or in the abstract. To do this, we simply replaced the [MeSH] tag with a [Title/Abstract] tag. Hence, the resulting V2 did not contain any explicit [MeSH] tags (Figure 1). Lastly, the V3 (MeSH-only) query was obtained by transforming all free-text terms into MeSH terms. It should be noted that terms stated as MeSH terms in the query but that do not

actually exist in the MeSH thesaurus are ignored by the PubMed engine; this is equivalent to deleting the terms (Figure 1).

The transformations from V1 to V2 and V3 were the same for all queries, regardless of whether they contained MeSH terms only or free-text terms only. However, we noted that some PubMed filters are based on MeSH terms [26]. It would therefore not be relevant to convert these terms into free-text terms. We drew up a list of these terms so that they were not transformed and were still able to serve as filters. Those 14 terms are the following: “80 and over”, “adolescent”, “adult”, “aged”, “animals”, “child”, “female”, “humans”, “infant”, “male”, “middle aged”, “newborn”, “preschool” and “young adult”.

Hence, each SR had a query written by the SR's authors (a combination of MeSH and free-text terms) (V1), a free-text-only query (V2), and an MeSH-only query (V3). Therefore, we intend to interpret the comparison of V2 to V1 as the added value of MeSH terms, and we intend to interpret the comparison of V3 to V1 as the added value of free-text terms.

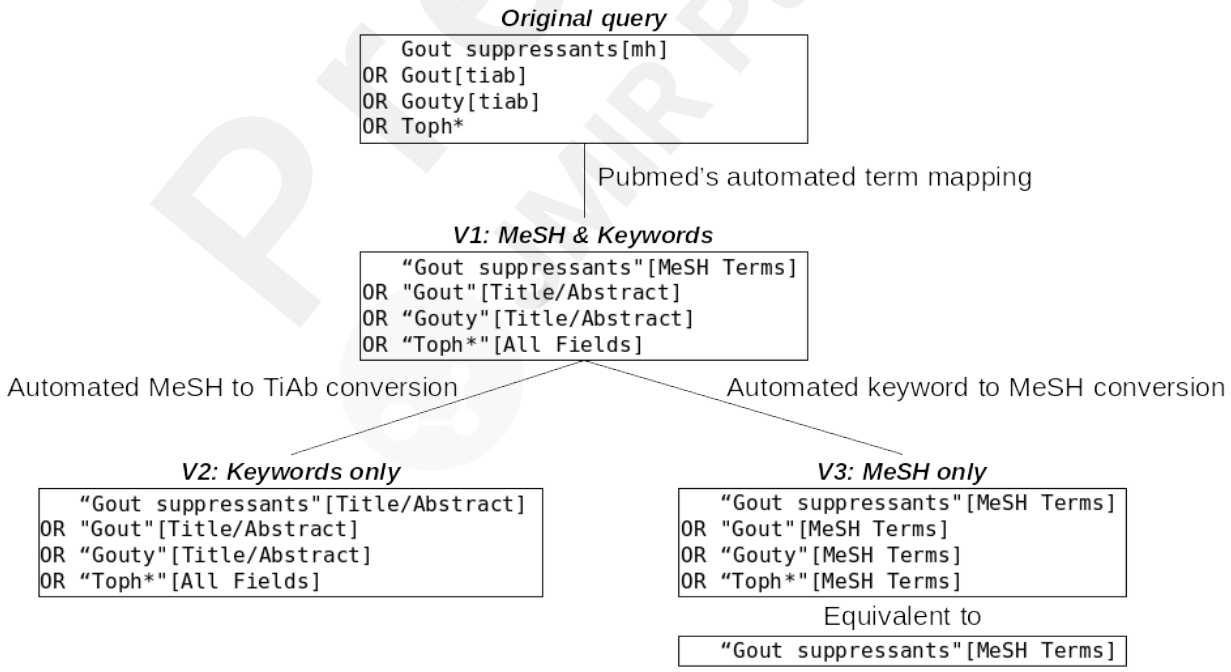


Figure 1. Example of an automated transformation of queries [27]

Each query was submitted to the PubMed search engine, and the results were retrieved and sorted by

the “Best Match” option. If there were more than 10,000 results, only the first 10,000 results were retained; in fact, PubMed does not allow more than 10,000 results to be extracted. The results were identified by their PubMed Identifier (PMID).

For each SR, the “gold standard” (GS) consisted of the articles selected by the authors of the systematic review. Each systematic review was read in order to extract the list of PMIDs selected by the authors. This work was done "by hand" by four researchers (VL, RB, BLG, AH). Publications cited in the SR but not indexed in MEDLINE were not analyzed. If the reference section did not contain the items selected in the SR, data extraction from supplementary files allowed for the completion of the gold standard.

Data analysis

For each SR, we obtained four lists of PMIDs: the GS, those retrieved by V1 (MeSH and free-text terms), those retrieved by V2 (free-text terms only), and those retrieved by V3 (MeSH terms only). For each list, we computed the sensitivity (also referred to as “recall”) and the positive predictive value (PPV, also referred to as “precision”) with respect to the GS. We then computed the F-score, which is the harmonic mean of the sensitivity and the PPV.

For each query *i* (V1, V2 and V3), the odds for the PPV was defined as the ratio between two numbers:

$$odds\left(PPV, query_i\right) = \frac{¿\left(query_i \cap GS\right)}{¿\left(query_i \cap \acute{G}S\right)}$$

Next, for a given SR and using the same GS, the odds ratio (OR) of $query_2$ to $query_1$ for the PPV was defined as:

$$¿\left(PPV, query_2, query_1\right) = \frac{odds\left(PPV, query_2\right)}{odds\left(PPV, query_1\right)}$$

Likewise, the odds for the sensitivity of each query *i* (V1, V2 and V3) was defined as the ratio between two numbers:

$$odds(Se, query_i) = \frac{¿(GS \cap query_i)}{¿(GS \cap qúery_i)}$$

Hence, for a given SR and using the same GS, the OR for $query_2$ vs. $query_1$ with regard to sensitivity was:

$$¿(Se, query_2, query_1) = \frac{odds(Se, query_2)}{odds(Se, query_1)}$$

We computed the respective ORs for V2 vs. V1 and V3 vs. V1 for the PPV and the sensitivity:

$¿(Se, query_2, query_1)$

$¿(Se, query_3, query_1)$

$¿(VPP, query_2, query_1)$

$¿(VPP, query_3, query_1)$

An OR of 1 means that the queries have the same level of performance with regard to the chosen indicator. An OR<1 denotes worse performance, and an OR>1 denotes better performance.

Statistical analysis

Qualitative variables, binary variables or discrete variables with very few modalities were expressed as the frequency (percentage). Quantitative variables were expressed as the mean (standard deviation (SD)) when symmetrically distributed and the median [interquartile range (IQR)] when not. The independence of two qualitative variables was probed in a chi-squared test.

All statistical tests were two-sided. The threshold for statistical significance was set to P<.05. The 95% confidence interval (CI) of a proportion was calculated using the Wald method. Statistical analyses were performed with R software, RStudio software, and the R ‘metafor’ package [28–30].

Regulatory framework and funding

This research did not receive any specific funding from agencies or organizations in the public, commercial, or not-for-profit sectors. The research was performed using publicly available

documents. It did not involve individuals or personal data. Approval by an institutional review board was not required.

Results

Flowchart

The SRs used to compile the set of queries were selected by a single researcher (Figure 2).

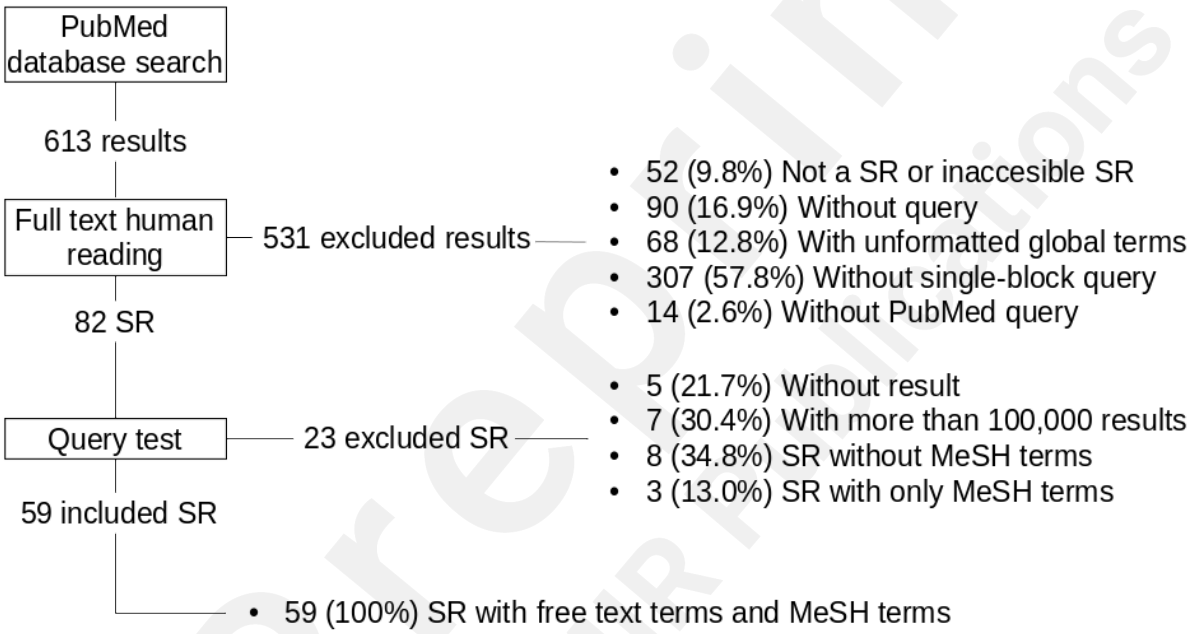


Figure 2 Flowchart for the selection of SRs

Description of the included SRs

A total of 59 SRs was selected for analysis, which contained both MeSH terms and free-text terms (Table 1 and Multimedia Appendix A & B).

Of the 59 selected SRs, 29 (49.2%) came from *The BMJ*, 19 (32.2%) came from the *Annals of Internal Medicine*, 6 (10.2%) came from *The Lancet*, and 5 (8.5%) came from the *Journal of the*

American Medical Association. The publication dates were evenly distributed: the mean publication year and the median publication year were both 2016.

The countries of origin of the first authors were available for 49 SRs (83.1%). The three most frequent countries of origin were the United States (n=21, 42.9%), the United Kingdom (n=5, 10.2%), and Canada (n=5, 10.2%).

Quantification of the utility of MeSH terms

The queries contained a median [IQR] of 43 [17.0; 98] terms. The median [IQR] number of MeSH terms in the V1 queries was 6.0 [3.0; 19.5]. The median [IQR] proportion of MeSH terms relative to all terms in queries was 18.5% [13.7; 25.5].

The V1 queries returned a total of 206,095 items, of which 1,628 (0.79%) were included in the GS (Table 1). The V2 queries returned a total of 157,698 items, of which 1,473 (0.93%) were included in the GS. In other words, an average of 820.29 additional articles per SR had to be screened for V1, relative to V2. Furthermore, V1 retrieved an average of 2.62 additional relevant articles, when compared with V2.

Table 1 General description of the attributes for each included SR
(GS = Gold Standard; Se = sensitivity; PPV = positive predictive value; F1 = F-score)

SR's PMID	Items Of V1	Items Of V2	Items Of V3	Items of GS	GS Ω V1	GS Ω V2	GS Ω V3	Se Of V1	PPV Of V1	F-Sc Of V1	Se Of V2	PPV Of V2	F-Sc Of V2	Se Of V3	PPV Of V3	F-Sc Of V3
33472813	297	294	147	30	27	27	24	0.900	0.091	0.165	0.900	0.092	0.167	0.800	0.163	0.271

33441384
4408
2265
901
16
14
13
11
0.875
0.003
0.006
0.812
0.006
0.011
0.688
0.012
0.024
33186535
145
153
1
6
4
4
1
0.667
0.028
0.053
0.667
0.026
0.050
0.167
1.000
0.286
33148618
9024
10000
3706
66
16
18
3
0.242
0.002
0.004
0.273
0.002
0.004
0.045
0.001
0.002
32909814
349
349
40
9
1
1
0
0.111
0.003
0.006
0.111
0.003
0.006
0.000
0.000
0.000
32496521
94

162
0
24
0
0
0
0.000
0.000
0.000
0.000
0.000
0.000
0.000
0.000
0.000
32459529
1950
812
1441
9
7
2
7
0.778
0.004
0.007
0.222
0.002
0.005
0.778
0.005
0.010
32442035
10000
10000
842
15
8
7
2
0.533
0.001
0.002
0.467
0.001
0.001
0.133
0.002
0.005
32371466
4185
3320
324
50
49
48
21
0.980
0.012
0.023
0.960
0.014
0.028
0.420
0.065
0.112
32199484
6347
6111
1231

128
0
0
0
0.000
0.000
0.000
0.000
0.000
0.000
0.000
0.000
0.000
31255301
1164
1023
456
61
41
35
34
0.672
0.035
0.067
0.574
0.034
0.065
0.557
0.075
0.132
30884526
770
716
30
79
0
0
0
0.000
0.000
0.000
0.000
0.000
0.000
0.000
0.000
0.000
30617123
1086
788
292
29
28
26
14
0.966
0.026
0.050
0.897
0.033
0.064
0.483
0.048
0.087
30326495
4360
2212
951
158
123

117
62
0.778
0.028
0.054
0.741
0.053
0.099
0.392
0.065
0.112
30158148
10000
2673
395
45
33
24
28
0.733
0.003
0.007
0.533
0.009
0.018
0.622
0.071
0.127
29049756
3155
2858
913
20
16
16
2
0.800
0.005
0.010
0.800
0.006
0.011
0.100
0.002
0.004
28903922
687
610
95
24
23
23
16
0.958
0.033
0.065
0.958
0.038
0.073
0.667
0.168
0.269
27893131
10000
10000
2439
48
34
33
27

0.708
0.003
0.007
0.688
0.003
0.007
0.562
0.011
0.022
27802505
2299
2227
195
21
19
19
14
0.905
0.008
0.016
0.905
0.009
0.017
0.667
0.072
0.130
27802478
3847
3525
2526
89
45
43
43
0.506
0.012
0.023
0.483
0.012
0.024
0.483
0.017
0.033
27548070
1634
910
626
26
20
10
9
0.769
0.012
0.024
0.385
0.011
0.021
0.346
0.014
0.028
27142267
10000
10000
10000
10
7
9
5
0.700
0.001

0.001
0.900
0.001
0.002
0.500
0.000
0.001
26903336
903
414
115
92
47
47
0
0.511
0.052
0.094
0.511
0.114
0.186
0.000
0.000
0.000
26349907
2675
265
1728
8
0
0
0
0.000
0.000
0.000
0.000
0.000
0.000
0.000
0.000
0.000
26199070
554
303
0
20
17
15
0
0.850
0.031
0.059
0.750
0.050
0.093
0.000
0.000
0.000
26109551
298
297
14
14
9
9
5
0.643
0.030
0.058
0.643

0.030
0.058
0.357
0.357
0.357
25770113
3046
2761
1956
7
7
5
7
1.000
0.002
0.005
0.714
0.002
0.004
1.000
0.004
0.007
25569206
746
691
0
49
49
49
0
1.000
0.066
0.123
1.000
0.071
0.132
0.000
0.000
0.000
25556126
212
206
50
9
8
8
4
0.889
0.038
0.072
0.889
0.039
0.074
0.444
0.080
0.136
25006006
834
395
407
25
24
23
20
0.960
0.029
0.056
0.920
0.058
0.110

0.800
0.049
0.093
24727842
2046
1989
0
69
66
66
0
0.957
0.032
0.062
0.957
0.033
0.064
0.000
0.000
0.000
24157497
10000
81
8636
61
61
0
61
1.000
0.006
0.012
0.000
0.000
0.000
1.000
0.007
0.014
24046285
978
978
812
12
12
12
12
1.000
0.012
0.024
1.000
0.012
0.024
1.000
0.015
0.029
23935058
628
20
887
5
3
0
1
0.600
0.005
0.009
0.000
0.000
0.000
0.200
0.001

0.002
23900314
499
300
195
6
5
5
4
0.833
0.010
0.020
0.833
0.017
0.033
0.667
0.021
0.040
23529983
283
160
0
8
8
2
0
1.000
0.028
0.055
0.250
0.013
0.024
0.000
0.000
0.000
23420235
6848
2305
3434
27
25
20
10
0.926
0.004
0.007
0.741
0.009
0.017
0.370
0.003
0.006
23033409
434
258
0
16
11
11
0
0.688
0.025
0.049
0.688
0.043
0.080
0.000
0.000
0.000
22986378

8633
7169
3884
40
26
38
19
0.650
0.003
0.006
0.950
0.005
0.011
0.475
0.005
0.010
22422870
1010
980
0
4
3
3
0
0.750
0.003
0.006
0.750
0.003
0.006
0.000
0.000
0.000
22323502
10000
10000
9784
5
3
3
2
0.600
0.000
0.001
0.600
0.000
0.001
0.400
0.000
0.000
22226047
1911
1239
1807
49
38
32
35
0.776
0.020
0.039
0.653
0.026
0.050
0.714
0.019
0.038
33176180
87
87

0
4
3
3
0
0.750
0.034
0.066
0.750
0.034
0.066
0.000
0.000
0.000
32479176
303
266
64
29
28
26
2
1.000
0.092
0.169
0.931
0.098
0.177
0.069
0.031
0.043
32427305
2035
2004
0
13
12
12
0
0.923
0.006
0.012
0.923
0.006
0.012
0.000
0.000
0.000
31727627
1884
1878
610
132
127
127
29
0.985
0.067
0.126
0.985
0.068
0.127
0.227
0.048
0.079
31585960
10000
10000
9514
227

186
206
139
0.819
0.019
0.036
0.907
0.021
0.040
0.612
0.015
0.029
30383109
4723
4716
570
38
36
36
2
0.947
0.008
0.015
0.947
0.008
0.015
0.053
0.004
0.007
28348110
27
27
0
24
1
1
0
0.042
0.037
0.039
0.042
0.037
0.039
0.000
0.000
0.000
28114600
85
35
25
68
9
3
0
0.132
0.106
0.118
0.044
0.086
0.058
0.000
0.000
0.000
26868137
10000
10000
10000
32
2
7

2
0.062
0.000
0.000
0.219
0.001
0.001
0.062
0.000
0.000
26830221
6167
1166
5102
76
74
28
72
0.974
0.012
0.024
0.368
0.024
0.045
0.947
0.014
0.028
26830055
6167
1166
5102
29
29
19
27
1.000
0.005
0.009
0.655
0.016
0.032
0.931
0.005
0.011
26420598
8405
8387
2421
57
47
47
30
0.825
0.006
0.011
0.825
0.006
0.011
0.526
0.012
0.024
26420387
8405
8387
2421
78
59
59
30
0.756

0.007
0.014
0.756
0.007
0.014
0.385
0.012
0.024
25059938
1524
1503
33
21
14
14
1
0.667
0.009
0.018
0.667
0.009
0.018
0.048
0.030
0.037
24592495
873
798
100
16
9
9
0
0.562
0.010
0.020
0.562
0.011
0.022
0.000
0.000
0.000
23460092
2426
2411
0
20
18
18
0
0.900
0.007
0.015
0.900
0.007
0.015
0.000
0.000
0.000
22777524
4645
3048
2838
37
37
35
31
1.000
0.008
0.016

0.946
0.011
0.023
0.838
0.011
0.022

Table 2 Comparison of the performance levels of queries V1, V2 and V3

	Query V1 (MeSH and FTTs)	*Query V2 (FTTs only)*	*Query V3 (MeSH only)*
Sensitivity	77.8 (62.1-95.2)	71.4 (42.6-90)	35.7 (0-61.7)
PPV	0.9 (0.3-2.8)	1.1 (0.3-3.4)	0.5 (0-2.6)
F-score	1.8 (0.7-5.4)	2.2 (0.7-6.1)	1 (0-3.9)
Number of results	1950 (657.50-6167.00)	1166 (301.50-2953.00)	456 (31.50-2188.50)
Number of gold standard (GS) items found	17 (7.00-36.50)	15 (4.50-32.50)	4 (0-22.50)
Number of results per GS item found	108.857 (35.062-298.574)	88.667 (29.682-314.848)	81.305 (20.99-564.125)

NB: all values are median (Q1 - Q3); FTTs = Free-text terms; n=59.

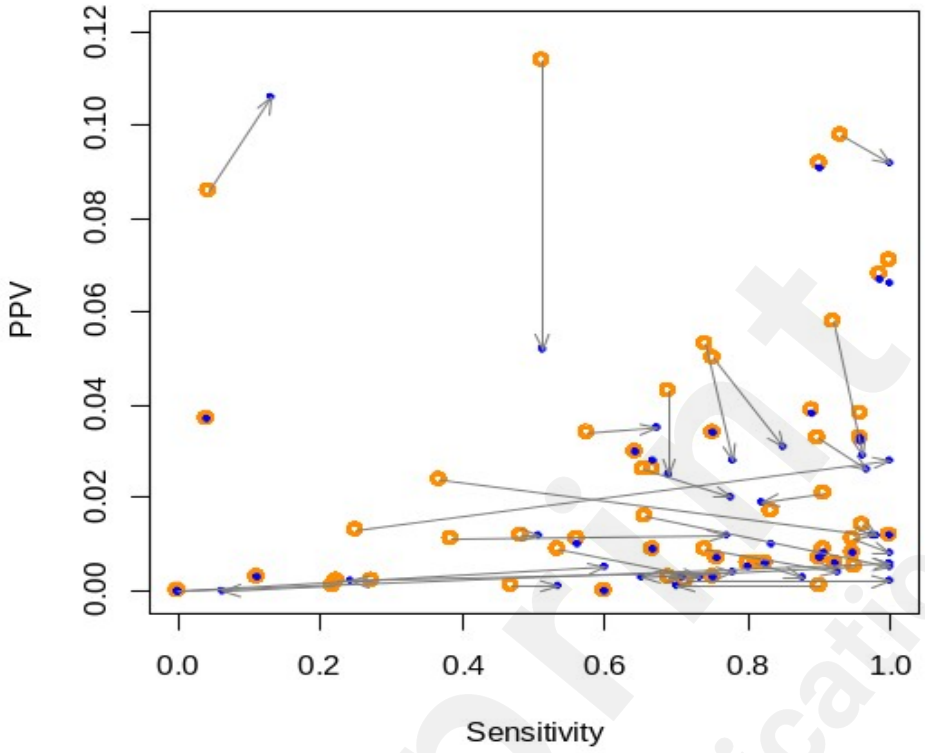


Figure 3 Contribution of MeSH terms to the queries. The orange circles correspond to V2 (free-text terms only), and the blue dots correspond to V1 (free-text terms and MeSH terms).

The medians (Q1-Q3) sensitivities of queries V1 and V2 were 77.8% (62.1%-95.2%) and 71.4% (42.6%-90.0%), respectively (Table 2). The medians (Q1-Q3) PPV of queries V1 and V2 were 0.9% (0.3%-2.8%) and 1.1% (0.3%-3.4%), respectively. The medians (Q1-Q3) F-scores of queries V1 and V2 were 1.8% (0.7%-5.4%) and 2.2% (0.7%-6.1%), respectively. A graphic visualization of the sensitivity and PPV per SR showed that the addition of MeSH terms to a query typically increased the sensitivity but decreased the PPV (Figure 3). Furthermore, it can be seen that the transition from V2 to V1 had no effect for many SRs.

Overall, V1 provided 8.49% more GS's items than V2 and 35.55% more GS's items than V3. V2 provided 27.06% more GS's items than V3. The ratio between the number of GS references retrieved by V1 and the number retrieved by V2 was within the interval [0, 1.05] in 39 (66.10%) cases (Figure 4). In 35 (59.32%) cases, the ratio was 1 or less. In other words, the transition from V1 to V2 did not have a marked effect on the number of relevant articles retrieved for more than half of the SRs.

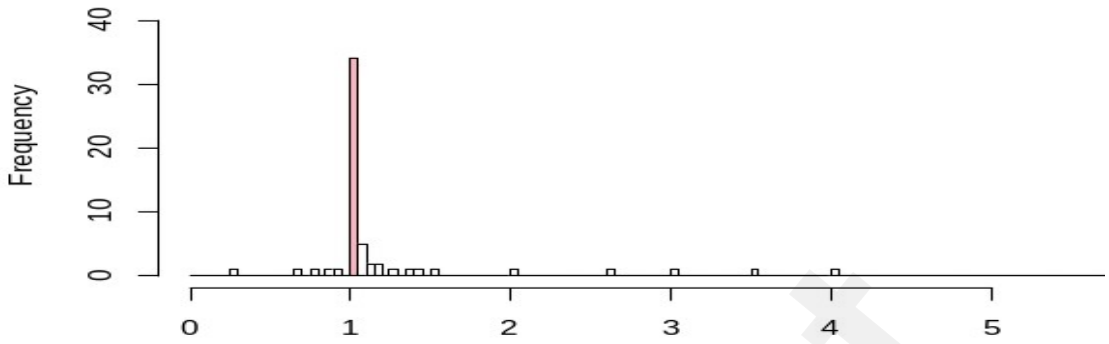


Figure 4 Distribution of the ratio between the number of relevant articles found by V1 and the number found by V2. The pink bar corresponds to the interval [1; 1.05]. For two cases, the ratio corresponded to the division of 0 by 0, and we considered that the result was 1. For other two cases, the result of the ratio corresponded to infinity (division by 0).

We also calculated the ORs for the number of relevant articles retrieved by V2 relative to V1 (Figure 5).

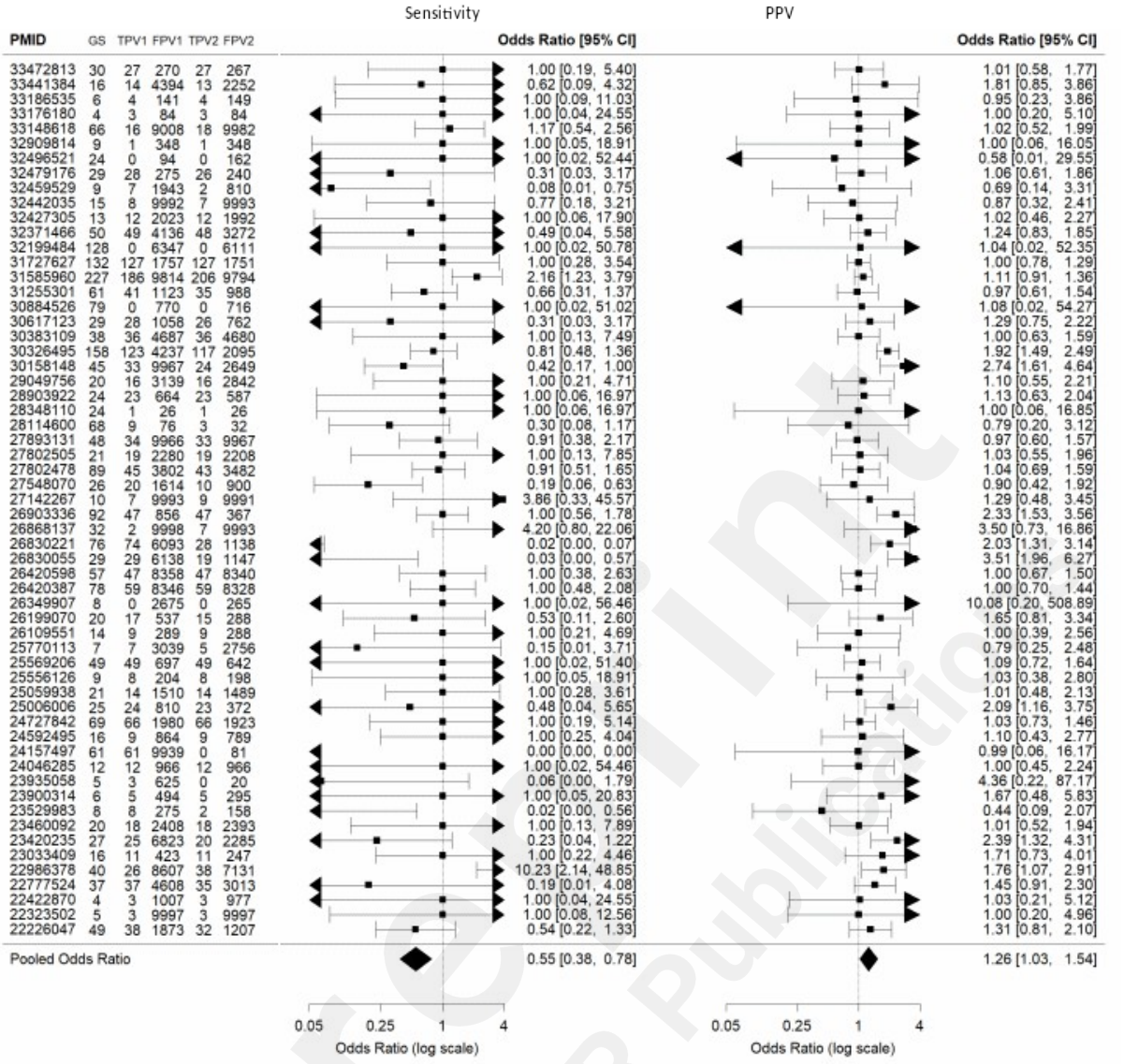


Figure 5 Forest plot of the OR for V2 vs. V1. An OR>1 means that V2 was better than V1 and so that inclusion of the MeSH terms was harmful. An OR<1 means that V2 was worse than V1 and so that MeSH terms were useful. (GS: Gold Standard, TPV1: True Positive V1, FPV1: False Positive V1, TPV2: True Positive V2, FPV2: False Positive V2)

Overall, the OR [95%CI] for V2 vs. V1 was 0.55 [0.38; 0.78] for sensitivity, 1.26 [1.03; 1.54] for the PPV (Figure 5). The OR [95%CI] for V3 vs. V1 was 0.31 [0.23; 0.41] for sensitivity, 3.11 [2.15; 4.48] for the PPV (Multimedia Appendix C).

Discussion

Key results

The objective of the present work was to quantify the utility of MeSH terms in SR queries. To this

end, we retrieved the queries drafted by the authors of 59 SRs published in four prestigious medical journals. We then modified the V1 query to give a free-text terms only query and a MeSH-only query. Lastly, we calculated the three queries' sensitivity, PPV, and F-score.

Our first key observation was that MeSH terms typically accounted for a non-negligible proportion (on average, 20.4%) of the terms in the query. Secondly, the removal of MeSH terms from SR queries decreased the sensitivity (by 6.4%, on the median), and increased the PPV (by 0.2%, on the median). In other words, queries containing both MeSH terms and free-text terms yield an average of 2.62 additional relevant papers per systematic review (SR), necessitating the screening of an additional 820.29 papers. The cost of screening an additional collected paper is therefore 313.09, which is slightly more than triple the mean reading cost associated with free-text terms only queries (88.67). Thirdly, our results indicate that the deletion of MeSH terms had no effect on the number of relevant articles retrieved for 35 of the 59 reviews (59.32%).

Discussion of the literature data

The results of a previous study were similar to those found here; 95% of the relevant articles were retrieved in 67% of the 73 analyzed SRs when the query contained free-text terms alone (relative to the V1 query with a mixture of MeSH terms and free-text terms) [22]. Another study with a similar objective gave significantly different results: the free-text terms only query was 25% less sensitive than MeSH-only [15]. However, it should be noted that (i) the latter findings were based on a single query, and (ii) the MeSH terms were converted to free-text terms manually, with a relatively limited set of synonyms used in the free-text terms strategy.

Three messages should be highlighted. First, MeSH terms remain an indispensable tool for systematic reviews despite the significant advancements in free-text search engines, especially in an era where the quality of systematic reviews is declining [31]. Second, free-text terms appear to contribute more effectively to the retrieval of relevant articles compared to MeSH terms. Third,

mixed queries (combining free-text and MeSH terms) exhibit poor positive predictive value; for rapid literature reviews, it is preferable to use either MeSH terms or free-text terms exclusively.

Our study involved queries developed by experienced researchers; choosing free-text terms can be challenging and requires expertise. It is possible that clinicians with limited experience in literature searching struggle to choose free-text terms effectively. And yet, bibliographic research among clinicians is essential [32]. MeSH terms offer a distinct advantage over free-text terms by covering a broad range of vocabulary, which can be particularly beneficial for, clinicians, early-career researchers or for non-native English speakers. In such cases, incorporating MeSH terms can help clinicians construct more comprehensive and effective queries.

Discussion of the method

The GS comprised solely MEDLINE-indexed documents with a PMID. This choice was restrictive but technically essential, given that the three queries were submitted to the PubMed search engine. However, our restriction to documents with a PMID increased the queries' sensitivity and decreased their PPV. We expect this bias to be non-differential, insofar as it should affect the three types of queries in the same way.

The publications with PMIDs 26420387 and 26420598 were written by the same authors and were based on the same search query. This was also the case for PMID 26830055 and 26830221. However, we considered these publications to be independent SRs, insofar as the corresponding GSs were different.

Strengths and weaknesses

Strengths

One strength of our study is that we used queries from a number of different researchers and research centers; this should mean that our results are more representative of currently used search strategies. Furthermore, the automatic transformation of V1 to V2 probably helped us to avoid bias associated

with differences in an individual’s knowledge of the MeSH thesaurus.

Weaknesses

Interpreting the results of V3 is delicate because the authors' queries are not designed to remain viable when ignoring all [tiab] & [all fields] etc. Indeed, after transformation to V3, 11 queries become non-viable and return zero items.

Additionally, it is important to note that the use of MeSH terms by the authors of the included SRs may be suboptimal and depends on each author's level of expertise. We assessed the quality of the MeSH selected by the authors of the included SR, not the actual utility of the MeSH as a feature. Finally, we are not able to measure the free-text terms retrieved from initial PubMed searches using only MeSH terms. However, the initial queries using MeSH terms alone may have enriched the search by helping to identify relevant free-text terms. It represents a potentially valuable contribution of MeSH terms that we do not measure here.

Perspectives

Our results and the literature data provide quantitive information on the use and value of MeSH terms in the queries used for SRs. MeSH terms still appear to be important for achieving a comprehensive SR. Our results also emphasized how difficult it is to build a query for an SR and highlighted the significant variability in the results obtained: the search strategies are a matter of concern for researchers [33–35]. With a view to gaining insights into the possible benefits of MeSH terms for use by less experienced researchers, it would be interesting to conduct a similar study of literature searches performed by clinicians. Finally, our study also highlights that any bibliographic research involves a tedious process of sifting through articles, akin to finding a needle in a haystack. While the authors of systematic reviews perform this task efficiently, inexperienced clinicians might find it discouraging to search for scientific articles. New tools based on network analysis [36] could help these clinicians find relevant articles more quickly.

Conclusion

The objective of this study was to estimate the utility of MeSH terms, selected by authors, in SR queries by analyzing the queries from 59 SRs published in four high-impact medical journals in general medicine. Our results revealed that removing MeSH terms from a query decreases sensitivity while slightly increasing the PPV. Queries containing both MeSH and free-text terms yielded more relevant articles but required screening many additional papers. Despite this additional workload, MeSH terms remain indispensable for systematic reviews and can be particularly beneficial for inexperienced clinicians or non-native English speakers, aiding in constructing more comprehensive queries. However, mixed queries combining MeSH and free-text terms show poor positive predictive value, suggesting the exclusive use of either MeSH terms or free-text terms for rapid reviews.

Conflicts of Interest

The authors have no conflicts of interest to declare.

References

1. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ 1996 Jan 13;312(7023):71–72. PMID:8555924

2. Bramer WM, de Jonge GB, Rethlefsen ML, Mast F, Kleijnen J. A systematic approach to searching: an efficient and complete method to develop literature searches. J Med Libr Assoc JMLA 2018 Oct;106(4):531–541. PMID:30271302

3. Brodman E, Field HG. Librarians' Symposia and Problems in Medical Subject Headings. Bull Med Libr Assoc 1947 Oct;35(4):287–295. PMID:16016776

4. Brodman E. Practical or Service Aspects of Medical Subject Headings. Bull Med Libr Assoc 1948 Apr;36(2):102–107. PMID:16016805

5. Larkey SV. Introduction to the Problem of Medical Subject Headings. Bull Med Libr Assoc 1948 Apr;36(2):69–81. PMID:16016801

6. Rogers FB. Medical subject headings. Bull Med Libr Assoc 1963 Jan;51(1):114–116. PMID:13982385

7. Doherty C, Joorabchi A, Megyesi P, Flynn A, Caulfield B. Physiotherapists' Use of Web-Based Information Resources to Fulfill Their Information Needs During a Theoretical Examination:

Randomized Crossover Trial. J Med Internet Res 2020 Dec 17;22(12):e19747. doi: 10.2196/19747

8. Lipscomb CE. Medical Subject Headings (MeSH). Bull Med Libr Assoc 2000 Jul;88(3):265–266. PMID:10928714

9. Rennesson M, Georget M, Paillard C, Perrin O, Pigeotte H, Tête C. Le thésaurus, un vocabulaire contrôlé pour parler le même langage. Médecine Palliat 2020 Feb;19(1):15–23. doi: 10.1016/j.medpal.2019.09.003

10. Salgado TM, Fernandez-Llimos F. Missing pharmacy-specific Medical Subject Headings (MeSH) terms: Problems and solutions. Res Soc Adm Pharm 2019 Sep;15(9):1189–1190. doi: 10.1016/j.sapharm.2019.04.008

11. Clarke M, Greaves L, James S. MeSH terms must be used in Medline searches. BMJ 1997 Apr 19;314(7088):1203–1203. doi: 10.1136/bmj.314.7088.1203a

12. Richter RR, Austin TM. Using MeSH (Medical Subject Headings) to Enhance PubMed Search Strategies for Evidence-Based Practice in Physical Therapy. Phys Ther 2012 Jan 1;92(1):124–132. doi: 10.2522/ptj.20100178

13. Klerings I, Robalino S, Booth A, Escobar-Liquitay CM, Sommer I, Gartlehner G, Devane D, Waffenschmidt S. Rapid reviews methods series: Guidance on literature search. BMJ Evid-Based Med 2023 Apr 19;bmjebm-2022-112079. doi: 10.1136/bmjebm-2022-112079

14. Jenuwine ES, Floyd JA. Comparison of Medical Subject Headings and text-word searches in MEDLINE to retrieve studies on sleep in healthy individuals. J Med Libr Assoc JMLA 2004 Jul;92(3):349–353. PMID:15243641

15. DeMars MM, Perruso C. MeSH and text-word search strategies: precision, recall, and their implications for library instruction. J Med Libr Assoc JMLA 2022 Jan 1;110(1):23–33. PMID:35210959

16. Dickersin K, Scherer R, Lefebvre C. Systematic Reviews: Identifying relevant studies for systematic reviews. BMJ 1994 Nov 12;309(6964):1286–1291. doi: 10.1136/bmj.309.6964.1286

17. Haynes RB, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC. Developing Optimal Search Strategies for Detecting Clinically Sound Studies in MEDLINE. J Am Med Inform Assoc 1994 Nov 1;1(6):447–458. doi: 10.1136/jamia.1994.95153434

18. Bachmann LM. Identifying Diagnostic Studies in MEDLINE: Reducing the Number Needed to Read. J Am Med Inform Assoc 2002 Nov 1;9(6):653–658. doi: 10.1197/jamia.M1124

19. Kassaï B, Sonié S, Shah NR, Boissel J-P. Literature search parameters marginally improved the pooled estimate accuracy for ultrasound in detecting deep venous thrombosis. J Clin Epidemiol 2006 Jul;59(7):710–714. doi: 10.1016/j.jclinepi.2005.09.013

20. Golder S, McIntosh HM, Duffy S, Glanville J. Developing efficient search strategies to identify reports of adverse effects in MEDLINE and EMBASE. Health Inf Libr J 2006 Mar;23(1):3–12. doi: 10.1111/j.1471-1842.2006.00634.x

21. Chang AA, Heskett KM, Davidson TM. Searching the literature using medical subject headings

versus text word with PubMed. The Laryngoscope 2006 Feb;116(2):336–340. PMID:16467730

22. Bramer WM, Giustini D, Kleijnen J, Franco OH. Searching Embase and MEDLINE by using only major descriptors or title and abstract fields: a prospective exploratory study. Syst Rev 2018 Nov 20;7(1):200. PMID:30458825

23. Clarivate. An Introduction to Journal Impact Factor. 2023. Available from: https://clarivate.com/products/scientific-and-academic-research/research-analytics-evaluation-and-management-solutions/journal-citation-reports/publishers/first-time-publishers/ [accessed Aug 24, 2023]

24. Garfield E. Journal impact factor: a brief review. CMAJ Can Med Assoc J J Assoc Medicale Can 1999 Oct 19;161(8):979–980. PMID:10551195

25. How PubMed works: Automatic Term Mapping (ATM). PubMed User Guide. 2023. Available from: https://pubmed.ncbi.nlm.nih.gov/help/#automatic-term-mapping [accessed Jun 1, 2023]

26. McKeever L, Nguyen V, Peterson SJ, Gomez-Perez S, Braunschweig C. Demystifying the Search Button: A Comprehensive PubMed Search Strategy for Performing an Exhaustive Literature Review. JPEN J Parenter Enteral Nutr 2015 Aug;39(6):622–635. PMID:26129895

27. Shekelle PG, Newberry SJ, FitzGerald JD, Motala A, O'Hanlon CE, Tariq A, Okunogbe A, Han D, Shanman R. Management of Gout: A Systematic Review in Support of an American College of Physicians Clinical Practice Guideline. Ann Intern Med 2017 Jan 3;166(1):37–51. PMID:27802478

28. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2022. Available from: https://www.R-project.org/

29. RStudio Team. RStudio: Integrated Development Environment for R. Boston, MA: RStudio, PBC.; 2020. Available from: http://www.rstudio.com/

30. Viechtbauer W. Conducting Meta-Analyses in *R* with the **metafor** Package. J Stat Softw 2010;36(3). doi: 10.18637/jss.v036.i03

31. Ioannidis JPA. The Mass Production of Redundant, Misleading, and Conflicted Systematic Reviews and Meta-analyses. Milbank Q 2016 Sep;94(3):485–514. PMID:27620683

32. McKibbon KA, Lokker C, Keepanasseril A, Wilczynski NL, Haynes RB. Net Improvement of Correct Answers to Therapy Questions After PubMed Searches: Pre/Post Comparison. J Med Internet Res 2013 Nov 8;15(11):e243. doi: 10.2196/jmir.2572

33. Lazarus JV, Palayew A, Rasmussen LN, Andersen TH, Nicholson J, Norgaard O. Searching PubMed to Retrieve Publications on the COVID-19 Pandemic: Comparative Analysis of Search Strings. J Med Internet Res 2020 Nov 26;22(11):e23449. PMID:33197230

34. Agoritsas T, Merglen A, Courvoisier DS, Combescure C, Garin N, Perrier A, Perneger TV. Sensitivity and predictive value of 15 PubMed search strategies to answer clinical questions rated against full systematic reviews. J Med Internet Res 2012 Jun 12;14(3):e85. PMID:22693047

35. Kastner M, Wilczynski NL, Walker-Dilks C, McKibbon KA, Haynes B. Age-specific search strategies for Medline. J Med Internet Res 2006 Oct 25;8(4):e25. PMID:17213044

36. BibliZap Team. BibliZap: an open-source and non-profit tool for reference mining that helps find similar articles. Lille, France. https://biblizap.org/

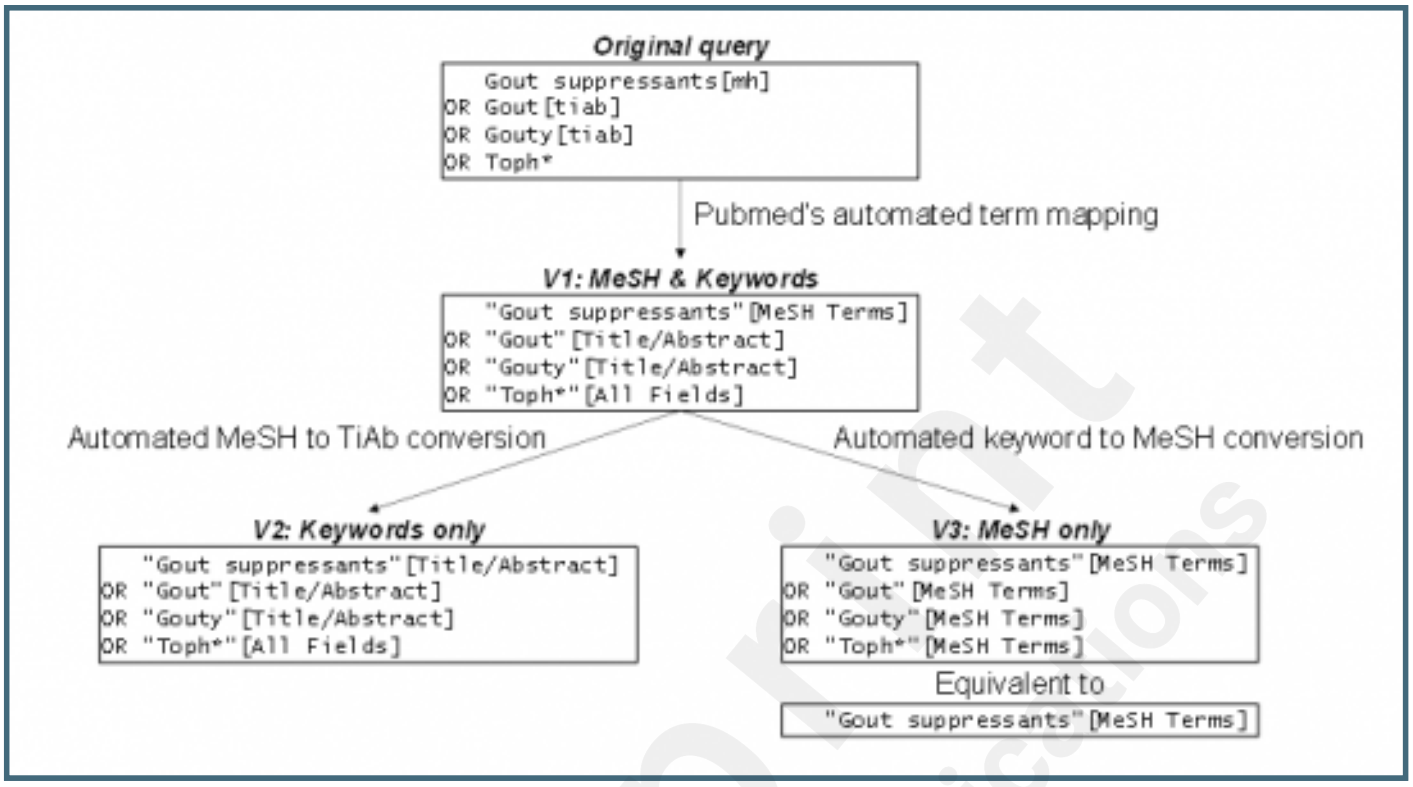
List of tables

List of figures

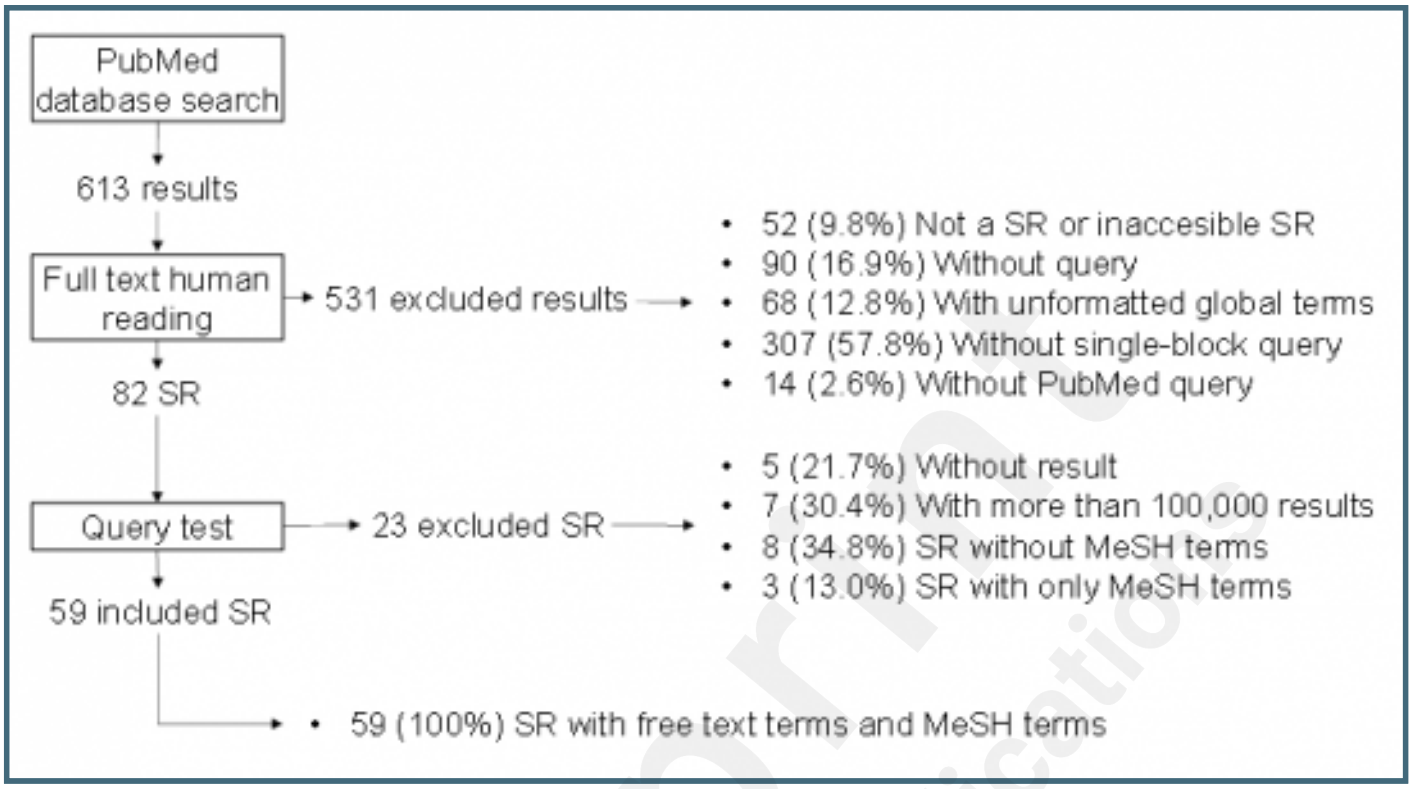
Supplementary Files

Figures

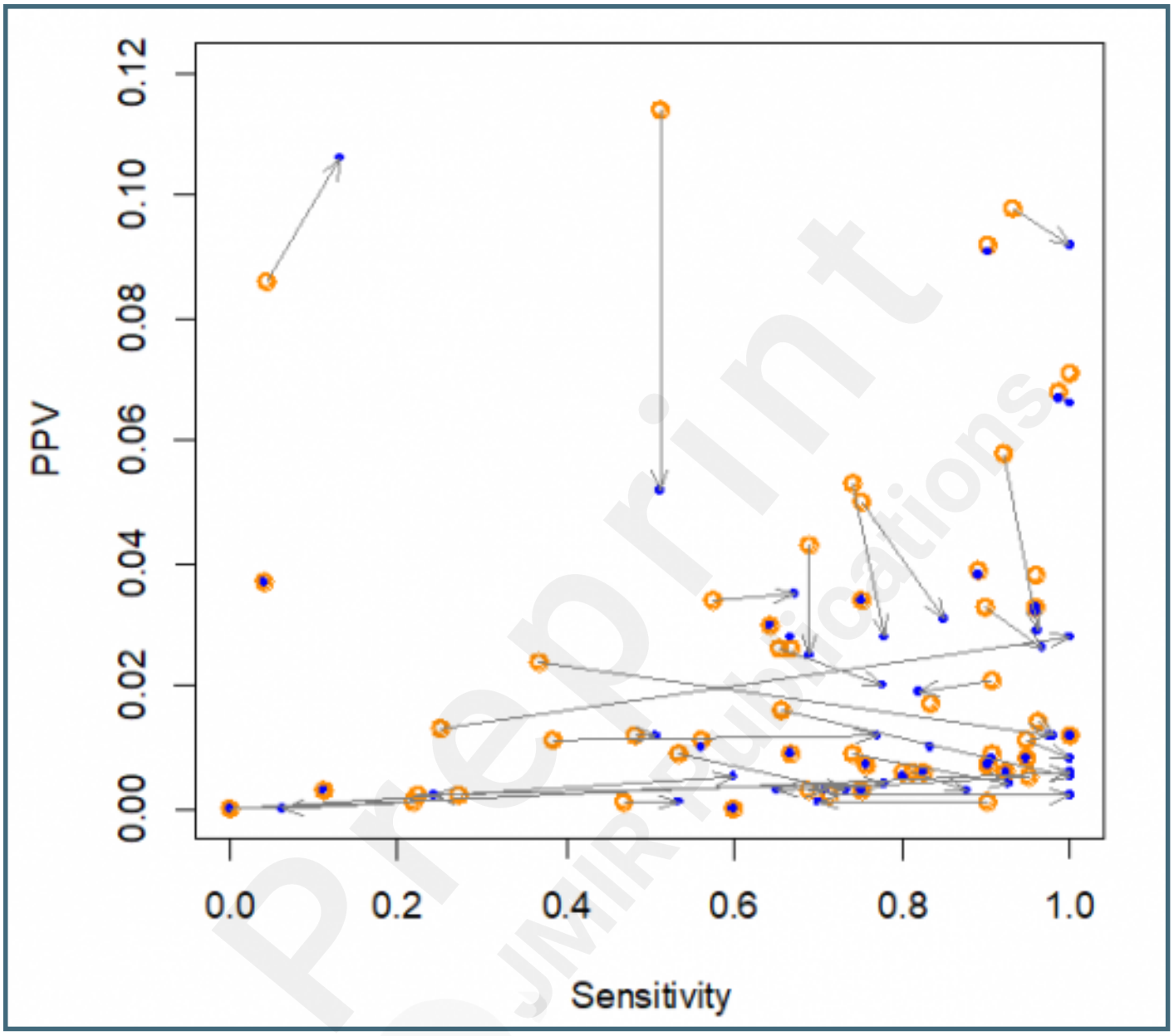
Example of an automated transformation of queries [27].



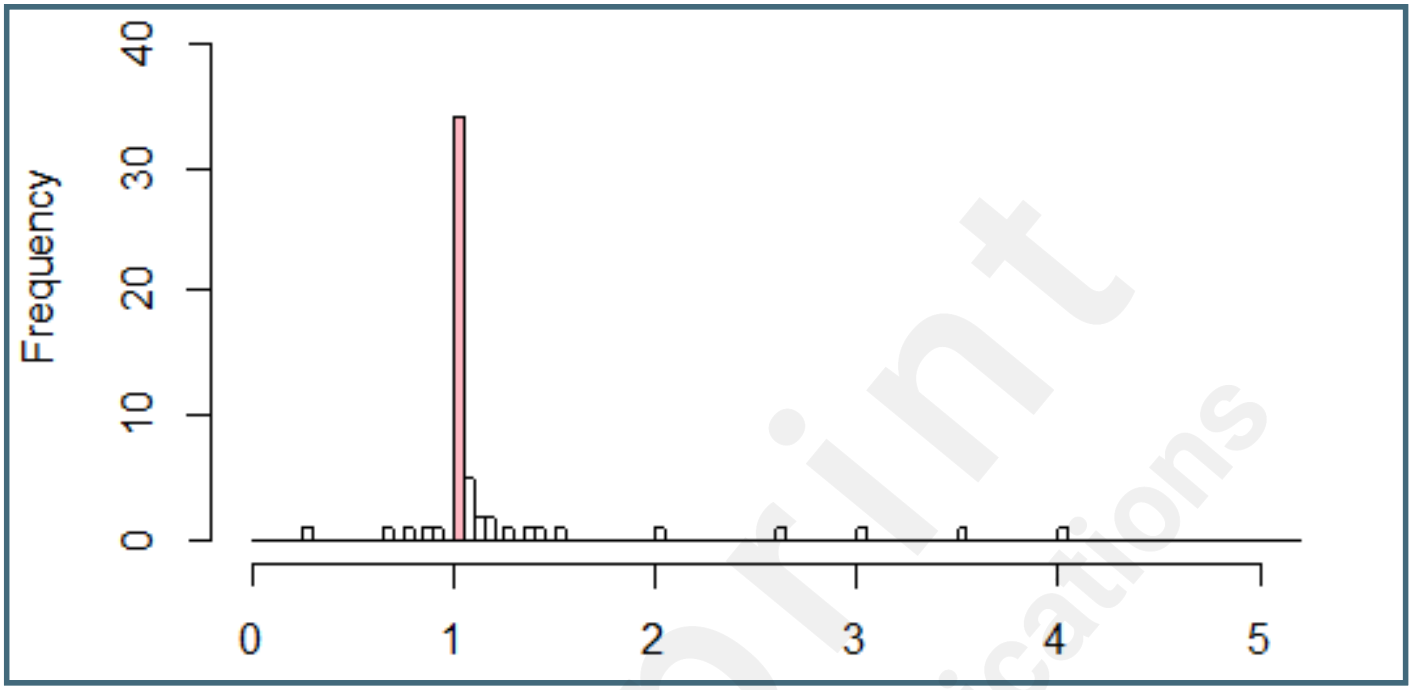
Flowchart for the selection of SRs.

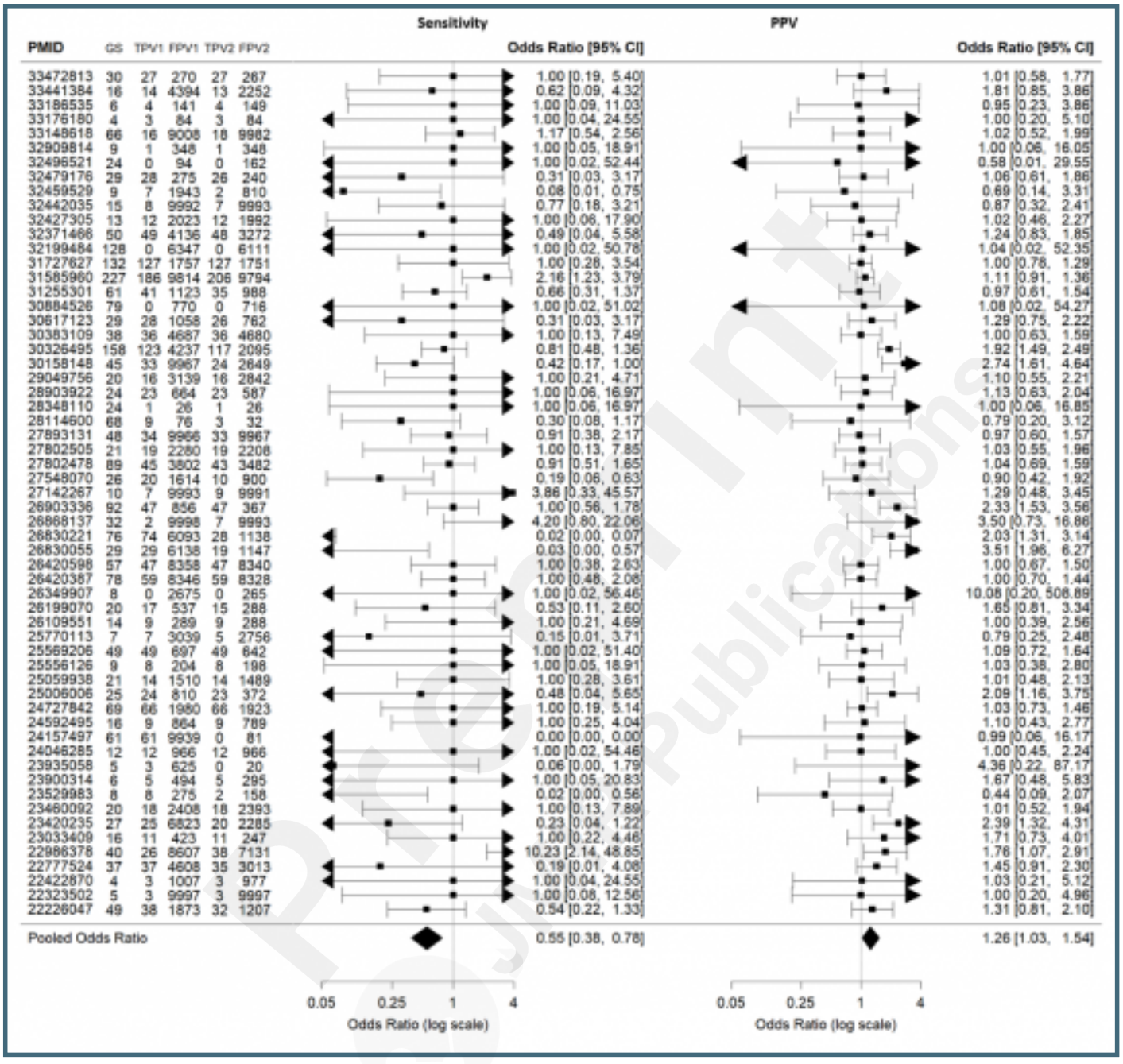


Contribution of MeSH terms to the queries. The orange circles correspond to V2 (free text terms only), and the blue dots correspond to V1 (free text terms and MeSH terms).



Distribution of the ratio between the number of relevant articles found by V1 and the number found by V2. The pink bar corresponds to the interval [1; 1.05]. For two cases, the ratio corresponded to the division of 0 by 0, and we considered that the result was 1. For other two cases, the result of the ratio corresponded to infinity (division by 0).




Sensitivity
PPV
PMID GS TPV1 FPV1 TPV2 FPV2
Odds Ratio [95% CI]
Odds Ratio [95% CI]
33472813 30 27 270 27 267 1.00 [0.19, 5.40] 1.01 [0.58, 1.77]
33441384 16 14 4394 13 2252 0.62 [0.09, 4.32] 1.81 [0.85, 3.86]
33186535 6 4 141 4 149 1.00 [0.09, 11.03] 0.95 [0.23, 3.86]
33176180 4 3 84 3 84 1.00 [0.04, 24.55] 1.00 [0.20, 5.10]
33148618 66 16 9008 18 9982 1.17 [0.54, 2.56] 1.02 [0.52, 1.99]
32909814 9 1 348 1 348 1.00 [0.05, 18.91] 1.00 [0.06, 16.05]
32496521 24 0 94 0 162 1.00 [0.02, 52.44] 0.58 [0.01, 29.55]
32479176 29 28 275 26 240 0.31 [0.03, 3.17] 1.06 [0.61, 1.86]
32459529 9 7 1943 2 810 0.08 [0.01, 0.75] 0.69 [0.14, 3.31]
32442035 15 8 9992 7 9993 0.77 [0.18, 3.21] 0.87 [0.32, 2.41]
32427305 13 12 2023 12 1992 1.00 [0.08, 17.90] 1.02 [0.46, 2.27]
32371466 50 49 4136 48 3272 0.49 [0.04, 5.58] 1.24 [0.83, 1.85]
32199484 128 0 6347 0 6111 1.00 [0.02, 50.78] 1.04 [0.02, 52.35]
31727627 132 127 1757 127 1751 1.00 [0.28, 3.54] 1.00 [0.78, 1.29]
31585960 227 186 9814 206 9794 2.16 [1.23, 3.79] 1.11 [0.91, 1.36]
31255301 61 41 1123 35 988 0.66 [0.31, 1.37] 0.97 [0.61, 1.54]
30884526 79 0 770 0 716 1.00 [0.02, 51.02] 1.08 [0.02, 54.27]
30617123 29 28 1058 26 762 0.31 [0.03, 3.17] 1.29 [0.75, 2.22]
30383109 38 36 4687 36 4680 1.00 [0.13, 7.49] 1.00 [0.63, 1.59]
30326495 158 123 4237 117 2095 0.81 [0.48, 1.36] 1.92 [1.49, 2.49]
30158148 45 33 9967 24 2649 0.42 [0.17, 1.00] 2.74 [1.61, 4.64]
29049756 20 16 3139 16 2842 1.00 [0.21, 4.71] 1.10 [0.55, 2.21]
28903922 24 23 664 23 587 1.00 [0.06, 16.97] 1.13 [0.63, 2.04]
28348110 24 1 26 1 26 1.00 [0.06, 16.97] 1.00 [0.06, 16.85]
28114600 68 9 76 3 32 0.30 [0.08, 1.17] 0.79 [0.20, 3.12]
27893131 48 34 9966 33 9967 0.91 [0.38, 2.17] 0.97 [0.60, 1.57]
27802505 21 19 2280 19 2206 1.00 [0.13, 7.85] 1.03 [0.55, 1.96]
27802478 89 45 3802 43 3482 0.91 [0.51, 1.65] 1.04 [0.69, 1.59]
27548070 26 20 1614 10 900 0.19 [0.06, 0.63] 0.90 [0.42, 1.92]
27142267 10 7 9993 9 9991 3.86 [0.33, 45.57] 1.29 [0.48, 3.45]
26903336 92 47 856 47 367 1.00 [0.56, 1.78] 2.33 [1.53, 3.56]
26868137 32 2 9998 7 9993 4.20 [0.80, 22.06] 3.50 [0.73, 16.86]
26830221 76 74 6093 28 1138 0.02 [0.00, 0.07] 2.03 [1.31, 3.14]
26830055 29 29 6138 19 1147 0.03 [0.00, 0.57] 3.51 [1.96, 6.27]
26420598 57 47 8358 47 8340 1.00 [0.38, 2.63] 1.00 [0.67, 1.50]
26420387 78 59 8346 59 8328 1.00 [0.48, 2.08] 1.00 [0.70, 1.44]
26349907 8 0 2675 0 265 1.00 [0.02, 56.46] 10.08 [0.20, 506.89]
26199070 20 17 537 15 288 0.53 [0.11, 2.60] 1.65 [0.81, 3.34]
26109551 14 9 289 9 288 1.00 [0.21, 4.69] 1.00 [0.39, 2.56]
25770113 7 7 3039 5 2756 0.15 [0.01, 3.71] 0.79 [0.25, 2.48]
25569206 49 49 697 49 642 1.00 [0.02, 51.40] 1.09 [0.72, 1.64]
25556126 9 8 204 8 198 1.00 [0.05, 18.91] 1.03 [0.38, 2.80]
25059938 21 14 1510 14 1489 1.00 [0.28, 3.61] 1.01 [0.48, 2.13]
25006006 25 24 810 23 372 0.48 [0.04, 5.65] 2.09 [1.16, 3.75]
24727842 69 66 1960 66 1923 1.00 [0.19, 5.14] 1.03 [0.73, 1.46]
24592495 16 9 864 9 789 1.00 [0.25, 4.04] 1.10 [0.43, 2.77]
24157497 61 61 9939 0 81 0.00 [0.00, 0.00] 0.99 [0.06, 16.17]
24046285 12 12 966 12 966 1.00 [0.02, 54.46] 1.00 [0.45, 2.24]
23935058 5 3 625 0 20 0.06 [0.00, 1.79] 4.36 [0.22, 87.17]
23900314 6 5 494 5 295 1.00 [0.05, 20.83] 1.67 [0.48, 5.83]
23529983 8 8 275 2 158 0.02 [0.00, 0.56] 0.44 [0.09, 2.07]
23460092 20 18 2408 18 2393 1.00 [0.13, 7.89] 1.01 [0.52, 1.94]
23420235 27 25 6823 20 2285 0.23 [0.04, 1.22] 2.39 [1.32, 4.31]
23033409 16 11 423 11 247 1.00 [0.22, 4.46] 1.71 [0.73, 4.01]
22986378 40 26 8607 38 7131 10.23 [2.14, 48.85] 1.76 [1.07, 2.91]
22777524 37 37 4608 35 3013 0.19 [0.01, 4.08] 1.45 [0.91, 2.30]
22422870 4 3 1007 3 977 1.00 [0.04, 24.55] 1.03 [0.21, 5.12]
22323502 5 3 9997 3 9997 1.00 [0.08, 12.56] 1.00 [0.20, 4.96]
22226047 49 38 1873 32 1207 0.54 [0.22, 1.33] 1.31 [0.81, 2.10]
Pooled Odds Ratio 0.55 [0.38, 0.78] 1.26 [1.03, 1.54]
0.05 0.25 1 4
Odds Ratio (log scale)
0.05 0.25 1 4
Odds Ratio (log scale)

Multimedia Appendixes

List of literature reviews analyzed.

URL: http://asset.jmir.pub/assets/4ff2f08c7426f0efc8231a6c5542701a.docx

All SR's queries: V1, V2, and V3.

URL: http://asset.jmir.pub/assets/20b1cfeb88e366553ae95dd75ee5da88.xlsx

Forest plot of the OR for V3 vs. V1. An OR>1 means that V3 was better than V1. (GS: Gold Standard, TPV1: True Positive V1, FPV1: False Positive V1, TPV3: True Positive V3, FPV3: False Positive V3).

URL: http://asset.jmir.pub/assets/0623034f149af5066668ca165025bbff.png

CONSORT (or other) checklists

URL: http://asset.jmir.pub/assets/e1a7c7b034212511a1bd6fdd3ecc0cb8.pdf