Received: 4 May 2020 | Revised: 29 August 2020 | Accepted: 4 September 2020

DOI: 10.1002/jsrm.1456

DISCUSSION PAPER/COMMENTARY

Supplement 1: The *Cited by N* link

D. V. Klopfenstein[1] | William Dampier*[2]

[1]School of Biomedical Engineering, Science, and Health Systems, Drexel University, PA, USA
[2]Department of Microbiology and immunology, Drexel University College of Medicine, PA, USA

Correspondence
*William Dampier. Email: wnd22@drexel.edu

Summary

First, we show the *Cited by N* links as they currently appear in PubMed and Google Scholar (GS). Second, we show the placement of the *Cited by N* links and the *N References* links that we would like to see in the PubMed Graphical User Interface (GUI). Finally, we demonstrate a new method to annotate PubMed search results with citation data from the NIH Open Citation Collection (NIH-OCC) for command line interface (CLI) users.

1 | THE CURRENT FORWARD CITATION SEARCH

Figures 1 and 2 show how to find both the citation count (Fig 2 red 3) and the reference count (Fig 2 red 4) in PubMed by searching for a paper using its title. Figure 3 shows how to find the citation count (Fig 3 red 2), but not the reference count, in Google Scholar (GS).

Clicking on the *See all "Cited by" articles* link (Fig 2, red 3a) in PubMed will open a new page showing the list of citations. This is a forward citation search. Clicking on the *Show all 35 references* link (bottom of Fig 2) will open a new page showing the 35 references. This a reverse citation search. Both forward and reverse citation search functionality are useful in exploratory searches and systematic searches. The reference count and reverse search functionality is not available in GS.

GS shows a total of 353 citations (Fig 3, bottom left red box) versus PubMed's 68 citation count. Since GS indexes preprints, white papers, theses, and papers not related to chemical and health sciences, it will show a higher citation count than PubMed. Higher citations counts do not always reflect better papers. For example, important research may have lower citation counts if it was published in a less prestigious journal. Also, citation count alone does not reflect how a paper is performing next to its peers. For example, a mathematical paper with a small number of citations may have a greater affect in its field than an oncology study having hundreds of citations[1].

2 | THE *CITED BY N* LINK

Clicking on the *Cited by 353* (Fig 3, bottom red box) in GS will open a page showing the first 10 or 20 citations (Fig 4). Researchers can not download the full 353 citations in bulk and instead must click and click and click to page through citation results 10 or 20 citations per page. However, GS's *Cited by N* links (Fig 4, red boxes) are exceedingly popular and are missing in PubMed. Citation count data can aid the researcher in deciding which paper to examine next.

Clicking on *See all "Cited by" articles* (Fig 2 red 3a text) link in PubMed will open a new page showing the list of citations (Fig 5). Researchers can download the full 68 citations, plus the original paper seen at the top, in a variety of formats including comma-separated values (csv), PubMed ID (PMID) lists, and summaries using the "SAVE" button boxed in red (Fig 5).

But the *Cited by N* link is missing. The additional links we would like to see in the PubMed interface are drawn in (Fig 6 red boxes). Note that only the top paper has "Cited by 108" because the following 3 papers are new and do not yet have citations.

In addition to the *Cited by N* link for easier forward search citation, we would like to see an *N References* link to aid in reverse citation search.

Notice that PubMed's "Best Match" relevance sorting shows new papers at the top of the citation list, while GS's relevance sorting shows highly cited papers at the top. GS favoring highly cited papers by ranking them at the top of their list[2], coupled with the laborious work of clicking and clicking to traversing the citation list can cause highly cited papers to continue to gain more citations, while important new papers may be overlooked.

3 | ATTACH CITATION DATA TO PUBMED RESULTS

After finding an article in PubMed, to obtain the equivalent functionality of having ubiquitously featured *Cited by N* links and *N References* links (Fig 6 red boxes) scroll down to the section showing all "Cited by" articles (bottom of Fig 1 and top of 2) and save all results as PMIDs to a file (Fig 7).

Next, the file can be annotated with citation and translational data from the NIH Open Citation Collection, iCite, using either the graphical user interface (GUI) method or the command line interface (CLI) method described below. These methods provide information equivalent to that found using GS's *Cited by N*. In this example, PubMed writes the list of PMIDs into a file named "pmid-26379270-set.txt."

3.1 | Using the Graphical User Interface (GUI)

If you are not a command-line interface (CLI) user, or you are a CLI user and want to see the iCite graphs generated for your data, go to the iCite website at https://icite.od.nih.gov/analysis and click on the "New Analysis" link at the top of the page. Load the file, in this case "pmid-26379270-set.txt", containing the your PMIDs (Fig 8) and press the "Process" bar. Click on the "Citations" tab (Fig 9 red 2), which is next to the tabs, "Influence" and "Translation." You will then see a list of the citing PMIDs, each annotated with citation count (Fig 9, red 3), reference count, and the Relative Citation Ratio (RCR), which is a NIH metric that compares the performance of a paper against its peers in the same citation network[3]. An RCR is a normalized metric of influence that has a transparent methodology and is free[3 1]. For example, a RCR of ~12 means that a paper is being cited 12 times as expected as compared to its peers.

3.2 | Using *pmidcite* from the Command Line Interface (CLI)

A Command-Line Interface (CLI) can be preferable to a Graphical User Interface (GUI) because: processing can be automated from a script; time-consuming mouse clicking is reduced; and more data can be seen at once on a text screen than in a browser, giving the researcher a better overall impression of the full set of information.

Linux and Mac users already work from the command line. Windows users can get that Linux-like command line feeling while still running native Windows programs by downloading Cygwin from https://www.cygwin.com/

3.2.1 | Run icite

The open-source software, *pmidcite*, is a Python package written by this first author and is freely available at https://github.com/dvklopfenstein/pmidcite. NIH's iCite has a simple, clean, fast Application Programming Interface (API) which allows their citation data to be downloaded over the internet by libraries like *pmidcite*. NIH's iCite returns a succinct record for each PMID containing citation, reference counts and more. When you run *pmidcite*, you are downloading the same data as seen in the previous GUI example. To run the "icite" script, which is part of the *pmidcite* library, do:

```
$ icite -i pmid-26379270-set.txt -o icite-26379270-set.txt
    69 WROTE: icite-26379270-set.txt
```

Running "`icite`" causes the PMIDs to be read from the PMID input ("`-i`") file specified with the argument "`-i pmid-26379270-set.txt`." The PMIDs are sent to NIH's iCite and citation data is downloaded and annotated to the PMIDs. The annotated PMID report is written to the output ("`-o`") file specified with the argument "`-o icite-26379270-set.txt`." The "`69 WROTE: icite-26379270-set.txt`" is text printed by icite letting the user know

that 69 PMIDs are annotated with citation data and written to "`icite-26379270-set.txt`." The PMIDs in the file are printed in the same order that they were sorted in PubMed with the "Best Match" algorithm before being downloaded.

3.2.2 | View citation data

The citation data is formatted in columns separated by spaces. To print the column headers and the citation data, two commands are run back-to-back. The first command is "`icite -H`", which causes the column headers to be printed. The semi-colon ("`;`") signifies the end first command. The second command is "`grep TOP icite-26379270-set.txt`." The command, "`grep`", is like the "find" command in word-processing programs, but it is used for ASCII text files. In this example, we are finding and printing all lines containing the text, "`TOP`", in the "`icite-26379270-set.txt`" citation report. The text, "`TOP`", signifies that the line contains citation data for a user-requested PMID, rather than a PMID that is a reference or citing paper of a user-requested PMID.

```
$ icite -H; grep TOP icite-26379270-set.txt
TYP PMID     RP HAMCc   % SD YR   cit cli ref au[00](authors) title
TOP 26379270 R. H....  99 4 2015   108  0  12 au[04](Neal Robert Haddaway) The Role of Google Scholar in E
TOP 32454855 .. H....  -1 i 2020     0  0  23 au[07](Huimin Jin) Effectiveness and Safety of Acupuncture M
TOP 32408896 R. H....  -1 i 2020     0  0  79 au[03](Emmanuel N-B Quarshie) Self-harm with suicidal and no
TOP 32235580 .. H....  -1 i 2020     0  0  40 au[02](Charlene Elliott) The Power of Packaging: A Scoping E
```

3.2.3 | Column headers

The column header line, which starts with "`TYP PMID ...`" is printed first by running "`icite -H`." Under the column header line is a list of papers, starting with the chosen Haddaway paper, and followed by papers that cite it. The columns, "`cit`" and "`ref`", contain the citation count and reference count for the paper on the same line. There is a high citation count of 108 for the Haddaway paper. We only show 3 of the 108 citing papers for brevity. The 2020 papers have a citation count of zero, likely due to being so recently published.

Notice that the citation counts for PubMed (68), NIH's iCite (108) and for GS (353) do not agree. The decision of which citation counts to use in the recommended new PubMed *Cited by N* links should be decided by PubMed. We do not believe that the differing citation counts are a major hindrance for a researcher in choosing their next paper to investigate while doing exploratory search. The data needed to choose the next paper to investigate is how new a paper is performing against its peers. Having the citations scaled lower, as in the PubMed citation counts will likely not have a great effect on displaying how each paper performs relative to others. Having the data to choose the next paper will speed the exploratory literature search faster than having all 353 GS citations. Once the researcher has become familiar with the subject through their exploratory literature search, then they can search in GS for additional papers.

The columns "`PMID`" and "`YR`" (YEAR) are the PubMed ID and the year that the paper was published.

3.2.4 | Column header key

To better understand the remaining columns containing additional NIH iCite data, print the key as shown below ("`icite -k`"). The key indicates that first the three letters of the line describes a paper's "type" (`TYP`), which include user-requested paper (`TOP`), citing paper (`CIT`), or reference (REF). The keyword makes grepping easier. Doing "`grep TOP [file.txt]`", will display one descriptive line for each user-requested PMID.

```
$ icite -k

KEYS TO PAPER LINE:
    TYP PubMedID RP HAMCc % nihSD YEAR x y z au[A](First Author) Title of paper

TYPe of relationship to the user-requested paper (TYP):
    TOP: A user-requested paper
    CIT: A paper that cited TOP
    CLI: A clinical paper that cited TOP
```

```
    REF: A paper referenced in the TOP paper's bibliography

NIH iCite details:

  PubMedID: PubMed ID (PMID)

     RP section:
     ----------------------------------
         R: Is a research article
         P: iCite has calculated an initial Relative Citation Ratio (RCR) for new papers

     HAMCc section:
     ----------------------------------
         H: Has MeSH terms in the human category
         A: Has MeSH terms in the animal category
         M: Has MeSH terms in the molecular/cellular biology category
         C: Is a clinical trial, study, or guideline
         c: Is cited by a clinical trial, study, or guideline

     NIH section, based on Relative Citation Ratio (RCR):
     ----------------------------------
         %: NIH citation percentile rounded to an integer. -1 means "not determined" or TBD
     nihSD: NIH citation percentile group: 0=-3SD 1=-2SD 2=+/-1SD 3=+2SD 4=+3SD or i=TBD

     YEAR/citations/references section:
     ----------------------------------
      YEAR: The year the article was published
         x: Number of unique articles that have cited the paper
         y: Number of unique clinical articles that have cited the paper
         z: Number of references
     au[A]: A is the number of authors
```

The sections, "RP" and "HAMCc" contain descriptive citation and translation data from the NIH-OCC and are explained in the key. The "y" number near the bottom of the key is the number of clinical trials that have cited the paper. If any clinical trials cite the paper, a "c", rather than a "." is printed in the "c" position of the "HAMCc" section.

3.2.5 | The NIH section

The key to the NIH section is:

```
     NIH section, based on Relative Citation Ratio (RCR):
     ----------------------------------
         %: NIH citation percentile rounded to an integer. -1 means "not determined" or TBD
     nihSD: NIH citation percentile group: 0=-3SD 1=-2SD 2=+/-1SD 3=+2SD 4=+3SD or i=TBD
```

The RCR is not used in the CLI display because the NIH developers calculated the NIH citation percentile using the RCR, making the NIH citation percentile a reasonable proxy for the RCR score. The NIH citation percentile is better suited to the CLI because it takes less space than the RCR and is better for sorting on the command line.

If the NIH percentile was displayed and used for sorting papers, the result would be unacceptably rigid sorting. Sorting such that a paper with a 55% NIH percentile is always shown ahead of a paper with a 45% percentile is not helpful, since both papers have a good citation performance. We created and added "NIH percentile group" ("nihSD"), so papers could be sorted by large performance groups, like good, high, and very high (Fig 10).

The performance groups are numbered from 0 to 4, with 0 being the lowest performing papers and 4 being the highest performing papers. Recent papers that have not had enough time to accumulate an RCR or NIH percentile are given an "i" as a placeholder for some number to be determined later.

3.2.6 | Summarizing a single paper

To print the summary for a single paper, do:

```
$ icite 27846867 -H
TYP PMID     RP HAMCc   % SD YR   cit cli ref au[00](authors) title
TOP 27846867 .. H....  82 2 2016    13  0  17 au[03](Claire Stansfield) Exploring issues in the conduct of
```

This paper has 13 citations ("cit") and 17 references ("ref"). Its NIH standard deviation group ("SD") is a 2, so it is a paper that is performing well with its peers. With an NIH percentile ("%") of 82%, it is almost in the high performing papers. The "PMID" for the paper is "27846867." The "-H" argument causes the column headers to be printed.

3.2.7 | Examining the citations of a single paper

To get all citation details for a single paper, do:

```
$ icite 27846867 --verbose --no_references
TOP 27846867 .. H....  78 2 2016    16  0  17 au[03](Claire Stansfield) Exploring issues in the conduct of
CIT 32511888 R. .....  -1 i 2020     0  0  32 au[03](Simon Briscoe) How do Cochrane authors conduct web se
CIT 31343759 .. H....  -1 i 2020     0  0  12 au[03](Louise Harriss) Building safety in humanitarian progr
CIT 31541534 .. H....  -1 i 2019     3  0  80 au[04](Anthea Sutton) Meeting the review family: exploring r
CIT 30993756 R. H....  -1 i 2019     2  0  13 au[11](Stefanie Buckner) Dementia Friendly Communities in En
CIT 29353363 RP H....  -1 i 2019     2  0   6 au[01](Marko Curkovic) The Implications of Using Internet Se
CIT 31866923 .. H....  -1 i 2019     0  0  51 au[07](Meg E Morris) Boxing for Parkinson's Disease: Has Imp
CIT 29783954 RP H....  98 4 2018    26  0 101 au[07](Monika Mueller) Methods to systematically review and
CIT 29193834 RP H....  90 3 2018     8  0  29 au[05](Chris Cooper) Supplementary search methods were more
CIT 29179733 .. H....  90 3 2017    10  0  49 au[04](Chris Cooper) A comparison of results of empirical st
CIT 29065246 .P H....  77 2 2018     7  0  51 au[01](Simon Briscoe) A review of the reporting of web searc
CIT 30177007 RP H...c  66 2 2018     5  2  40 au[05](Filippo Bianchi) Restructuring physical micro-environ
CIT 30424741 RP H....  66 2 2018     5  0  16 au[02](Marko Ćurković) Bubble effect: including internet sea
CIT 30340498 RP HA..c  42 2 2018     3  2  44 au[05](Filippo Bianchi) Interventions targeting conscious de
CIT 30453942 RP H....  17 2 2018     1  0  58 au[05](N Mahmoodi) Are publicly available internet resources
CIT 29137503 R. H....  42 2 2017     4  0   6 au[11](Yasser Bhatti) Global Lessons In Frugal Innovation To
CIT 28789703 .. H....   0 0 2017     0  0  37 au[02](Rosie Hanneke) Information sources for obesity preven
```

The "--verbose" option causes the citing papers to be printed.
The "--no_references" option prevents the references from being printed.

The requested paper, PMID "27846867" appearing on the "TOP" line, is printed along with its citing papers. Because the citations were downloaded using NIH-OCC and not from PubMed, PubMed's "Best Match" relevance sort has not been run.

So *pmidcite*'s default sort of the citations ("CIT") begins with sorting the group numbers in this order: "i, 4, 3, 2, 0." Second, it sorts by year so that the newest papers are at the top of each group. Third, it sorts by NIH percentile ("%").

Since all the newest papers do not have a NIH percentile yet ("-1"), the NIH percentile sort has no effect. For all new papers in the same year, the next sort uses the sum of the citations from both clinical ("cli") and non-clinical ("cit") papers. Finally, for new papers published in the same year that have the same number of citations, the final sort is done using the number of references ("ref").

To get all citation details for the references for a single paper do:

```
$ icite 27846867 --verbose | grep REF
REF 26379270 R. H....  99 4 2015   119  0  12 au[04](Neal Robert Haddaway) The Role of Google Scholar in
REF 26052848 R. .....  98 4 2014    86  0   8 au[03](Quenby Mahood) Searching for grey literature for syst
REF 27686611 R. H....  94 3 2016    24  0  37 au[07](Jean Adams) Searching and synthesising 'grey literatu
REF 26932789 R. H....  91 3 2016    29  0  11 au[03](Wichor M Bramer) Comparing the coverage, recall, and
REF 26494010 R. H....  97 3 2015    49  0  13 au[05](Katelyn Godin) Applying systematic review search meth
REF 24360284 R. H....  86 3 2013    39  0  24 au[04](Wichor M Bramer) The comparative recall of Google Sch
REF 25928625 R. .....  61 2 2015    14  0  13 au[04](Sandy Oliver) Capacity for conducting systematic revi
REF 25889619 R. H....  61 2 2015    10  0   9 au[01](Simon Briscoe) Web searching for systematic reviews:
REF 25031558 R. H....  76 2 2014    29  0  23 au[04](Julie M Glanville) Searching ClinicalTrials.gov and t
REF 24785398 .. H...c  66 2 2014    18  1  45 au[05](Rebecca W Rees) 'It's on your conscience all the time
REF 26052650 R. .....  51 2 2014    11  0  11 au[05](Tamara Rader) Methods for documenting systematic revi
REF 26052653 R. .....  44 2 2014     9  0  13 au[03](Claire Stansfield) Search wide, dig deep: literature
REF 24160679 R. H....  80 2 2013    31  0  40 au[03](Martin Boeker) Google Scholar as replacement for syst
REF 21439062 .. H...c  79 2 2011    35  2  24 au[04](Rebecca Rees) The views of young children in the UK a
REF 11706930 R. .....  40 2 2001    18  0   0 au[03](G Eysenbach) Evaluation of the usefulness of Internet
REF 23738438 R. H....  13 1 2013     3  0   0 au[01](Karen Blakeman) Finding research information on the w
REF 26061784 R. H....   8 1 2011     2  0  19 au[02](Karen Schucan Bird) Systematic literature searching i
```

The "`grep REF`" command causes only the references to be shown on the screen, not the user-requested ("`TOP`") paper or its citing papers ("`CIT`"). The default sort order of the references is the same as the default sort order for the citations.

To sort the most recent papers by reference count ("`ref`"), do:

```
$ icite 27846867 --verbose | grep -w i | sort -k10,10 -r
CIT 31541534 .. H....  -1 i 2019     3  0  80 au[04](Anthea Sutton) Meeting the review family: exploring r
CIT 28789703 .. H....  -1 i 2017     0  0  36 au[02](Rosie Hanneke) Information sources for obesity prever
CIT 30993756 R. H....  -1 i 2019     2  0  13 au[11](Stefanie Buckner) Dementia Friendly Communities in En
CIT 29353363 RP H....  -1 i 2019     2  0   6 au[01](Marko Curkovic) The Implications of Using Internet Se
```

To get only the most recent papers that have not yet been rated, "`grep -w i`" command looks for the lone "` i `", in the report lines, which represents the NIH SD data. The "`-w`" option on grep means "`i`" must be a "word" by itself, not embedded in another word. Another grep that would produce the same result is "`grep "-1 i"`."

This list is sorted by the number of references. So the paper with the most references ("80") is at the top and the paper with the smallest number of references is at the bottom ("6"). The "`-k10,10`" sort option causes the sort to only use the 10th column which is the reference count ("`ref`"). If you used "`-k10`", the sort would start from the text at the 10th column and ending at the end of the line. In this case, since all of the reference counts are different there is no difference in the sort when using "`-k10`" or "`-k10,10`". Because "`-k10`" is shorter to type, it may preferable to use in this circumstance.

3.2.8 | Retaining a history

To retain a history of papers of interest, re-run "`icite`", appending the paper's NIH iCite information to a log file ("`-a lit.txt`"). If the name of the log file is "`lit.txt`", do:

```
$ icite 27846867 -a lit.txt
```

The "`lit.txt`" can be saved and revision-managed using a tool like git and can be a reminder as to which papers were of interest and the order they were found.

4 | COMMENT ON *N REFERENCES* LINK

Researchers can access a document's references using PubMed, but not GS. Only being able to traverse the literature in the forward direction by traversing to new papers that cited a paper, but not being able traverse backwards to references in the paper hinders a researcher's ability to dig deep into the literature.

Worse, it prevents a single author from confirming that a paper is erroneously marked as *Cited by* by another paper in GS. This potential breach in literature connectivity in GS may result in incorrect citations and the inability to measure the extent of this potential problem.

5 | CONCLUSION

The search experience in PubMed is superior to GS in many ways. But GS is good secondary source for researchers who want to find unreviewed, unpublished grey literature, which includes material like white papers, government documents, and annual reports. GS's "*N*" in its *Cited by N* will be higher than PubMed's "*N*" due to its inclusion of such materials.

We urge biomedical scientists who now use GS as their number one search interface for academic literature searches to take a look at PubMed and consider it for their primary source and GS for their secondary source.

References

1. Naik G. The quiet rise of the NIH's hot new metric. *Nature* 2016; 539(7628): 150–150. doi: 10.1038/539150a

2. Mayr P, Walter AK. An exploratory study of Google Scholar. *Online information review* 2007; 31(6): 814–830.

3. Hutchins BI, Yuan X, Anderson JM, Santangelo GM. Relative Citation Ratio (RCR): A New Metric That Uses Citation Rates to Measure Influence at the Article Level. *PLOS Biology* 2016; 14(9): e1002541. doi: 10.1371/journal.pbio.1002541

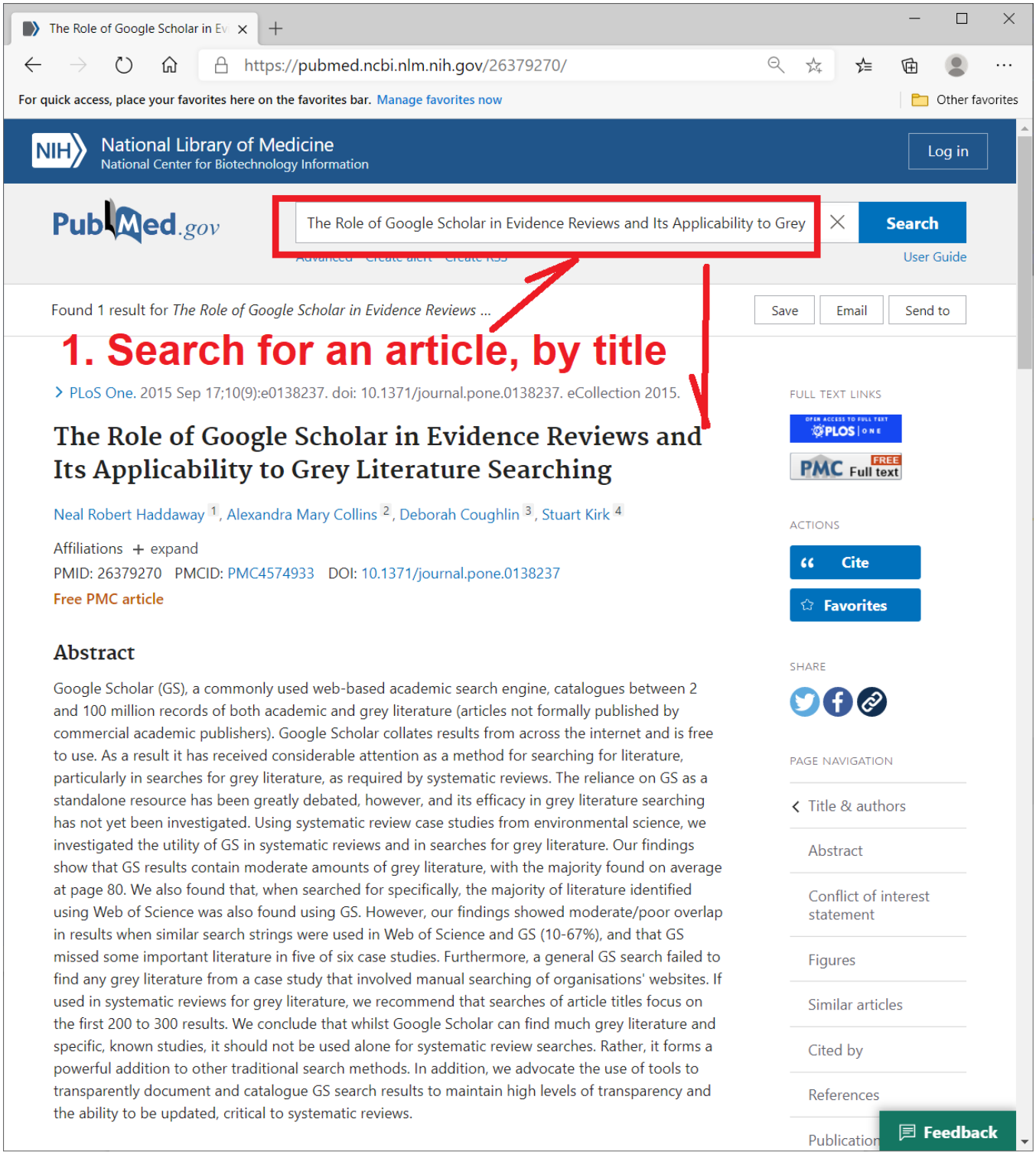

FIGURE 1 Searching for a specific paper by it's title in PubMed. First, enter the article's title in the search box. Then press enter. Scroll down to see the forward citation.

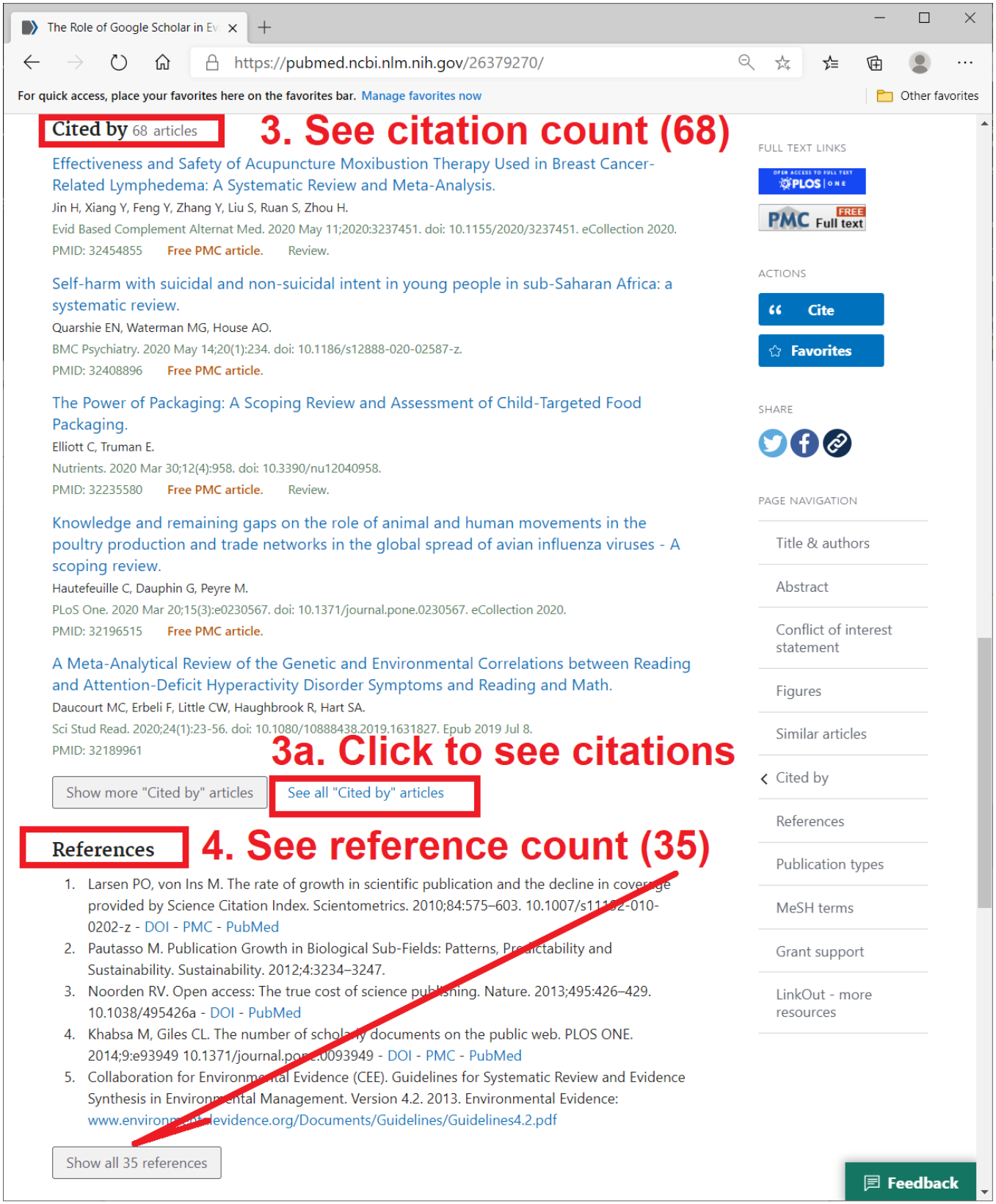

FIGURE 2 Searching for a specific paper by it's title in PubMed. After entering the article's title in the search box and pressing enter, scroll down to see the forward citation in the "Cited by" section. The header of the "Cited by" section and the "References" section contain the count of citing items. To open a page with the list of citations, click *See all "Cited by" articles*, marked as step 3a.

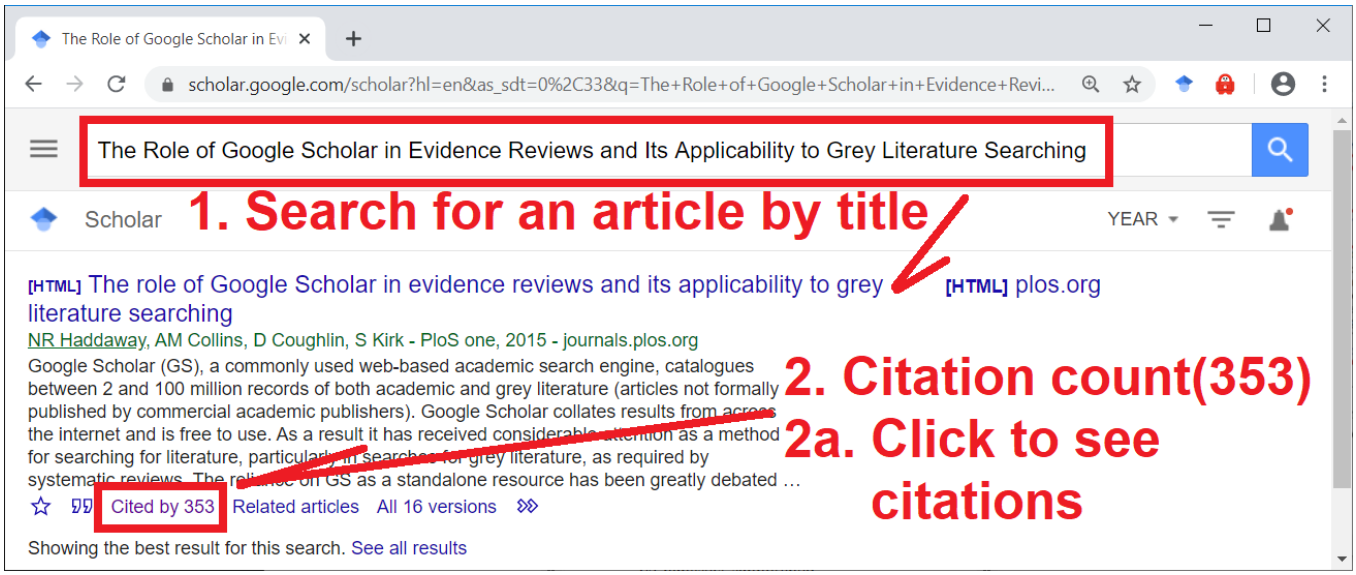

FIGURE 3 Searching for a specific paper by it's title shows 353 citations in Google Scholar. First, enter the article's title in the search box. Then press enter. Then see the citation count.

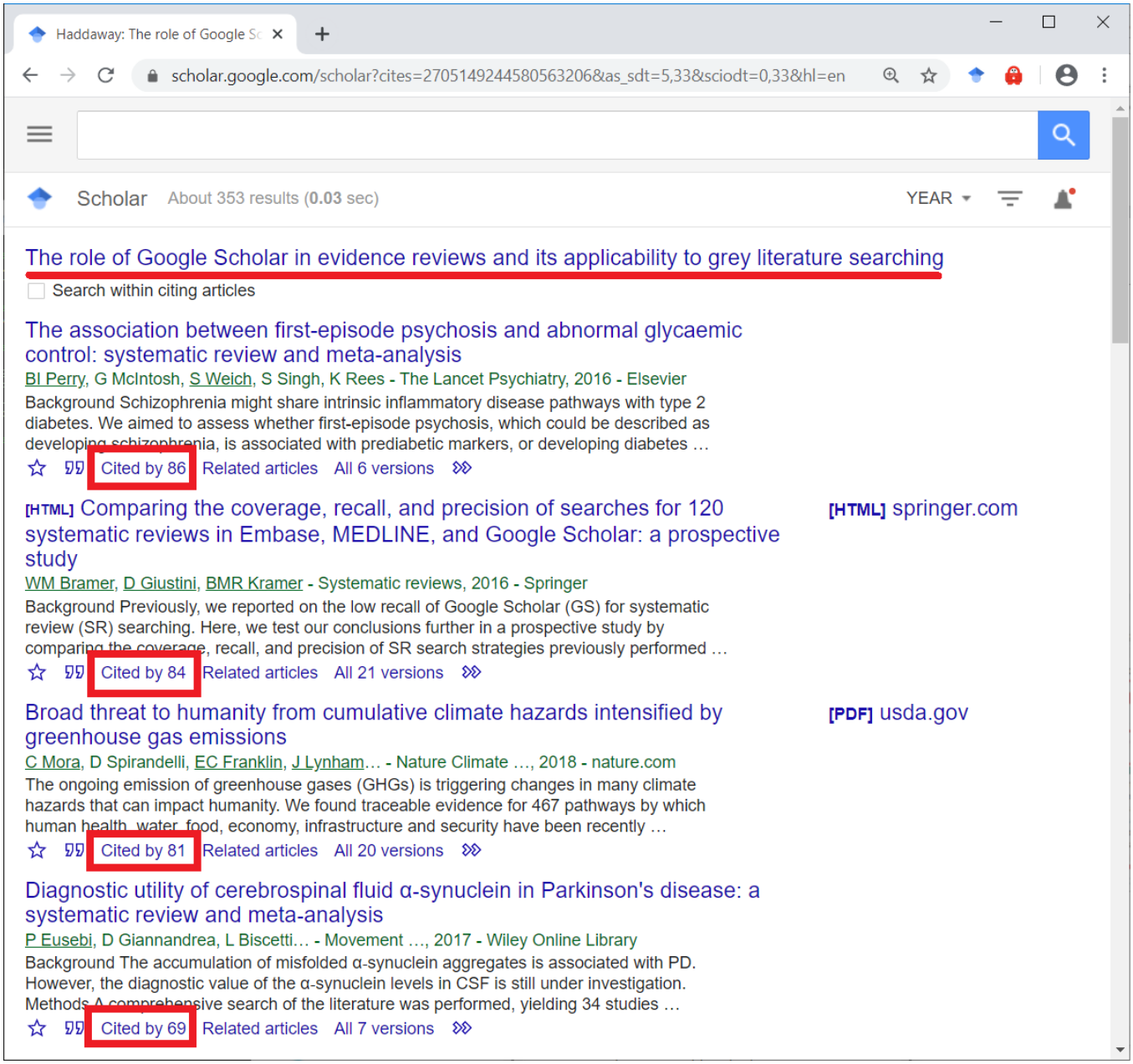

FIGURE 4 The first four of 353 citations for a specific paper. Each citation for the paper, "The Role of Google Scholar in Evidence Reviews and Its Applicability to Grey Literature Searching" has its own citation count, shown in the red boxes, aiding the user to choose the next paper to investigate.

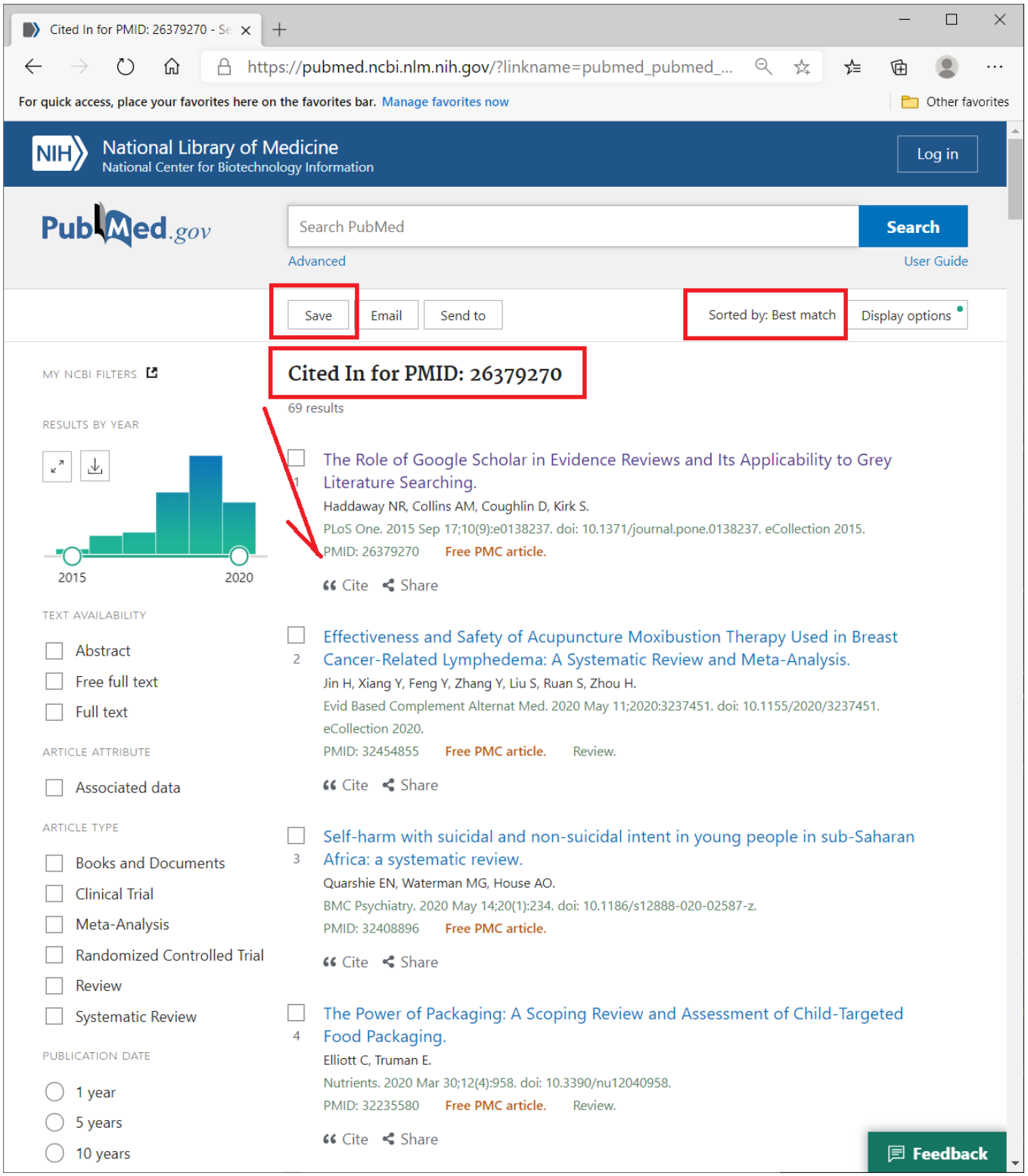

FIGURE 5 The first three of 68 citations for a specific paper as they currently appear in PubMed. The original paper, "The Role of Google Scholar in Evidence Reviews and Its Applicability to Grey Literature Searching", is at the top of the list. All 68 citing papers can be downloaded or shown on the page by scrolling down and clicking, "Show more."

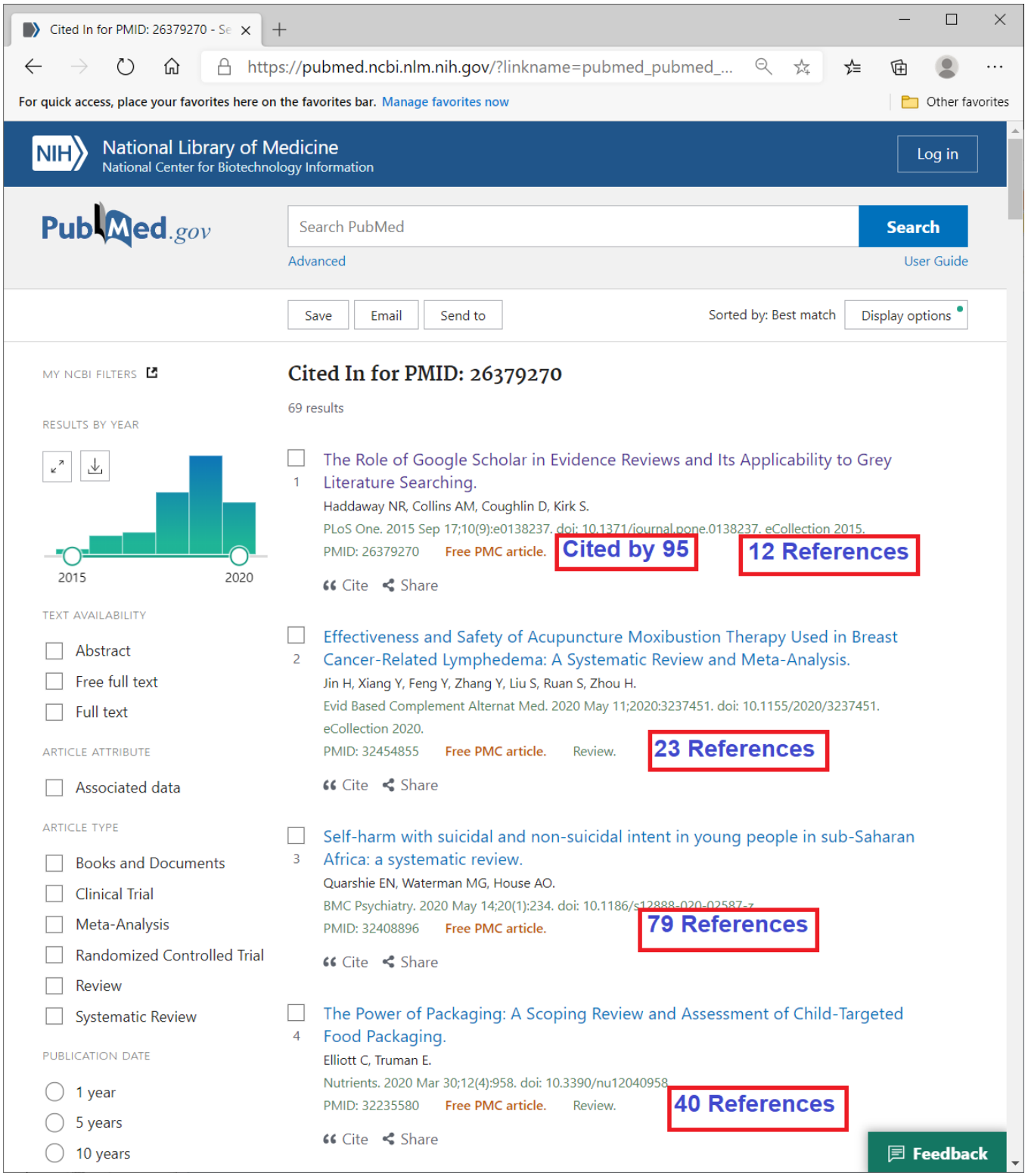

FIGURE 6 The first three of 68 citations for a specific paper as we would like it to appear in PubMed. We would like to see a *Cited by N* link and a *N References* link on the same line as the PubMed ID (PMID). Recently published citing papers (items 2 through 4) lack a *Cited by N* link because they do not yet have citations which are associated with a PMID.

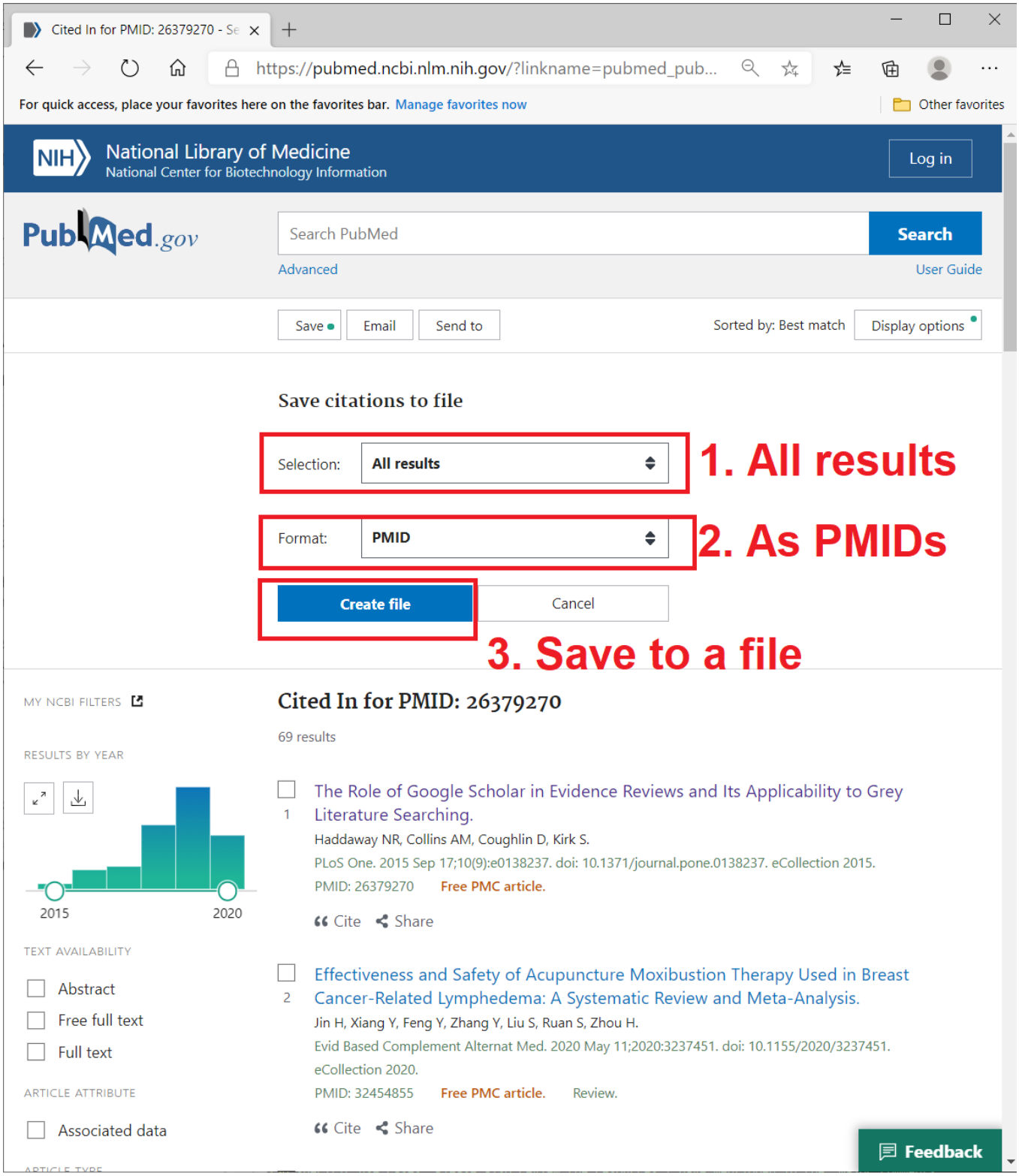

FIGURE 7 Save PMIDs into a file. Click the "Save" button. Then save all results as PMIDs into a file.

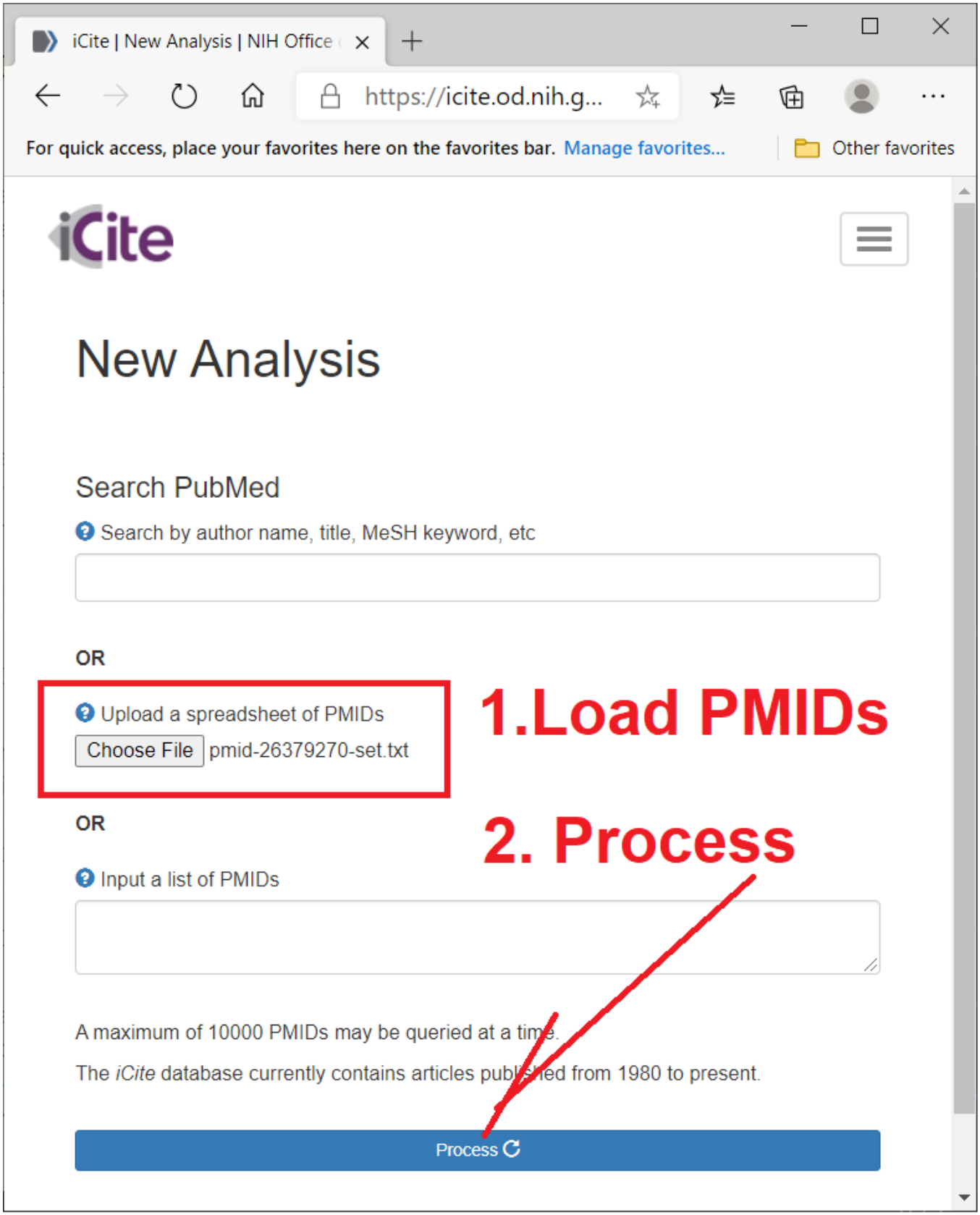

FIGURE 8 Annotate PMIDs with NIH iCite citation data. Load the PMIDs into iCite and process.

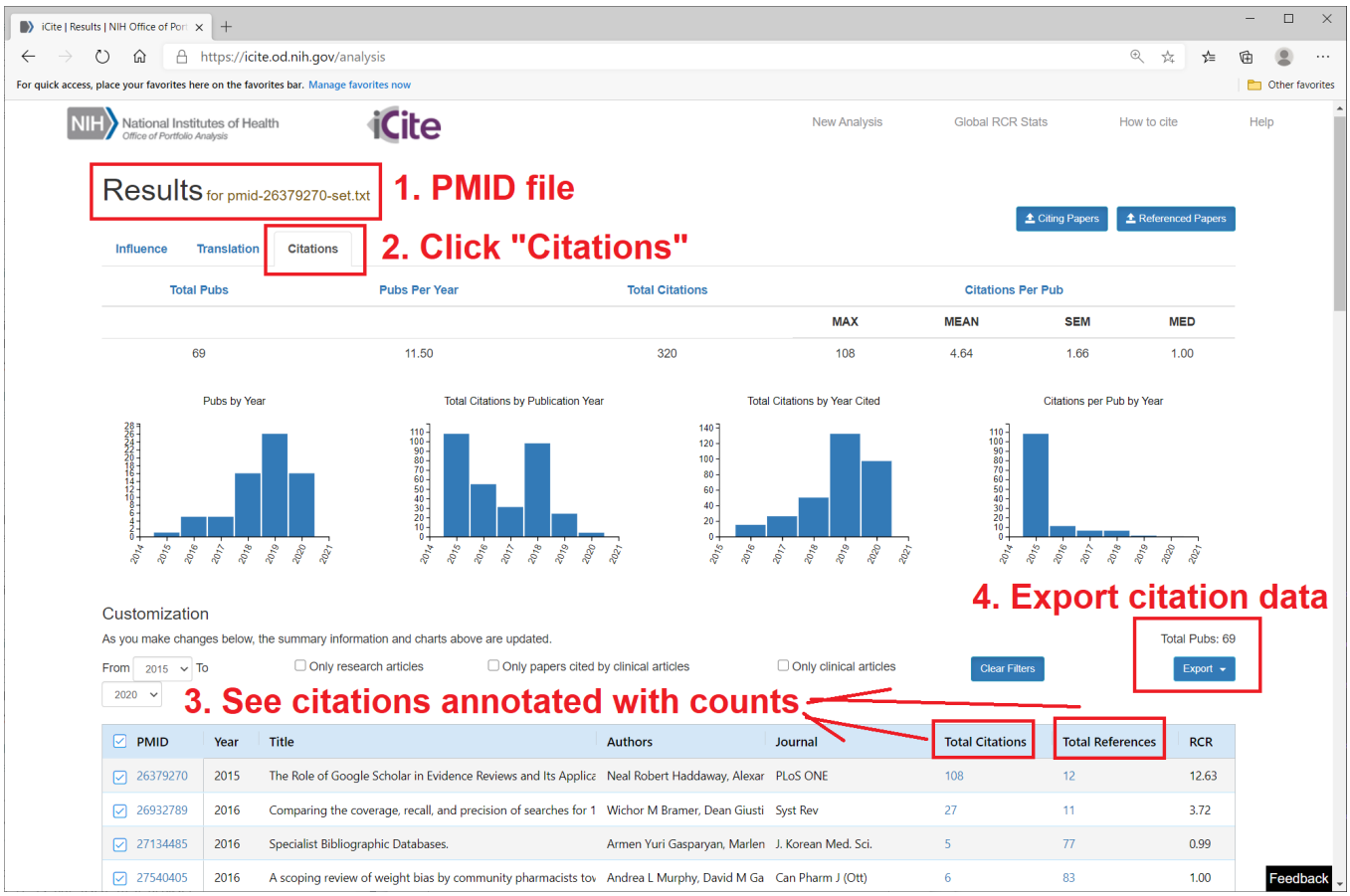

FIGURE 9 Annotate PMIDs with NIH iCite citation count data. Annotated citations appear as a list under the red "3" and do not retain their order sorted by PubMed's "Best Match" algorithm. The annotated citation data can be exported (red "4").

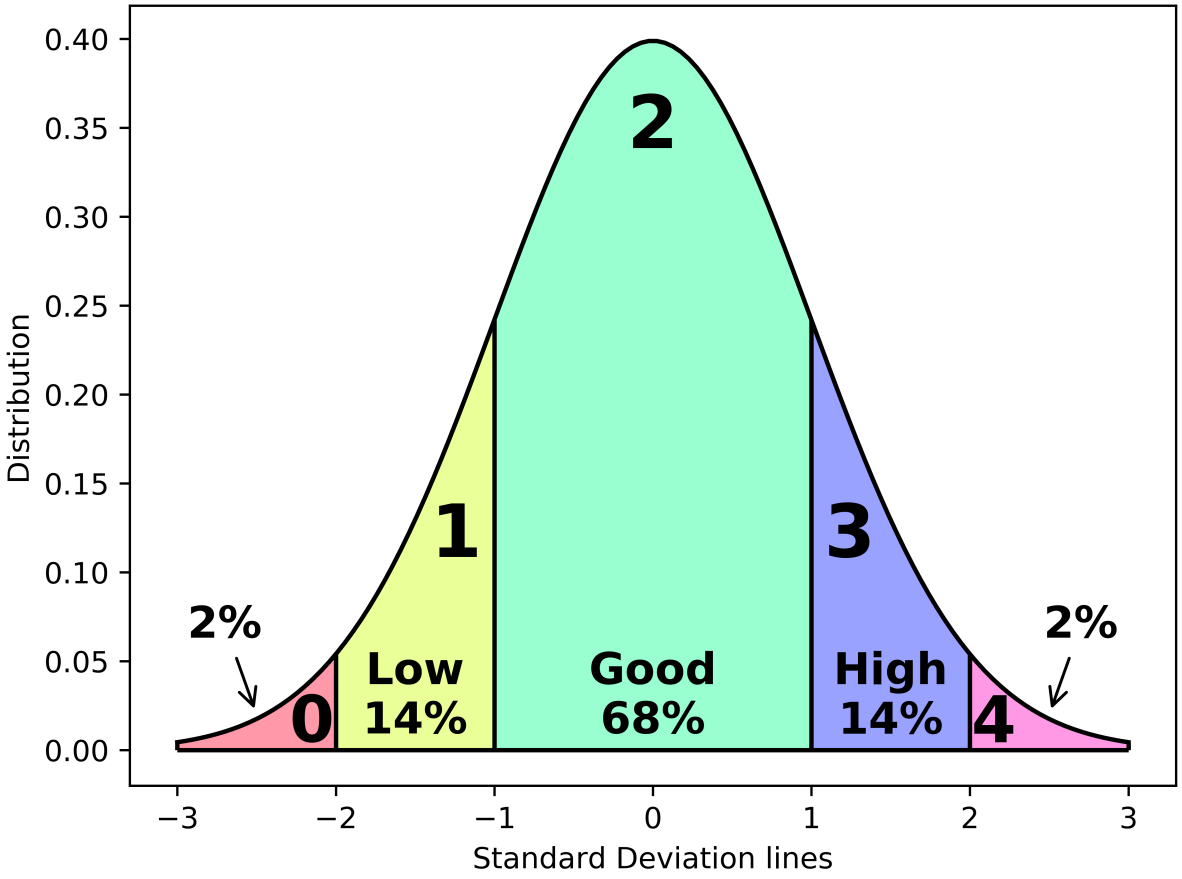

FIGURE 10 NIH percentile groups. There are 5 numbered NIH percentile groups. The lowest performing papers are given a 0. The highest performing papers are given a 4.

How to cite this article: D. V. Klopfenstein and W. Dampier (2020), Commentary to Gusenbauer and Haddaway 2020: Evaluating retrieval qualities of Google Scholar and PubMed, *Research Synthesis Methods*, *20NN;00:N–N*.