Merge search results from Pubmed and Web of Science

Siyue Yang, Jessica Gronsbell

06/15/2022

We extract articles in PubMed and Web of Science on the following journals/conferences:

- Journal of American Medical Informatics Association (JAMIA)
- JAMIA Open
- Journal of Biomedical Informatics (JBI)
- PloS One
- Proceedings of the Annual American Medical Informatics Association Symposium (AMIA)

Here are the results returned on April 14, 2022.

Table 1: Number of articles extracted by search queries

Source	n
PubMed	745
Web of Science	651
Total	1396

We followed the procedure below to identify duplicates and merge the search results:

Overview of the merging procedure

1. We extracted 745 articles from PubMed, from which we removed 28 AMIA articles accepated in 2017, resulting a total of 717 PubMed articles.
2. We extracted 651 articles from Web of Science, and removed 4 duplication (grouping by PMID), resulting a total of 647 Web of Science articles. We also corrected the journal name error for 1 article.
3. We merged the datasets of 717 PubMed articles and 647 Web of science articles, from which we identified 516 duplicates. Within the duplicated articles, We also identified 2 papers with their correction, and we removed the corrected version.
4. There are total 850 articles passed the filters.

In the subsequent sections, we describe the details and reasons:

1. Publication time conflict in PubMed

We noticed that there is a gap between 2018-2020 in PubMed extracted AMIA articles, as shown in Table 2.

Table 2: Number of AMIA articles extracted by PubMed

Publication.Year	n
2018	58
2020	20
2021	21
2022	14

We copied and pasted the titles in Google Scholar and found that there may be a gap between acceptance date and publish date of AMIA articles. PubMed is likely to extract AMIA articles using the acceptance date but record the publish date in the csv file. We also checked the AMIA PubMed journal list and validated our assumptions. The records of time conflict do not occur in Web of Science.

In order to merge the results from two database with minimum error, we manually searched the titles in Google Scholar and corrected the year of publication for these AMIA articles (in `amia20220414.csv`). We then remove all AMIA articles accepted in 2017.

Table 3: Number of AMIA articles extracted by PubMed (we will remove all articles in 2017)

Publication.Year	n
2017	28
2018	30
2019	20
2020	21
2021	14

2. Incorrect record information in Web of Science

There is an article "Extraction of Active Medications and Adherence Using Natural Language Processing for Glaucoma Patients" accepted by AMIA 2022 is recorded with a different journal/conference title as "OHSU Digital Commons". We mannually changed it as "AMIA" in the next section when merging.

Two articles extracted by Web of Science did not have a PMID. They are "Sleep apnea phenotyping and relationship to disease in a large clinical biobank" and "Generating real-world data from health records: design of a patient-centric study in multiple sclerosis using a commercial health records platform". We added PMIDs for them manually.

Additionally, the data extracted by Web of Science contains two articles with each occurs 3 times and we removed the 4 duplicated versions.

3. Merge articles from the two database

We follow the following steps to merge articles from PubMed and Web of Sciences. The implementation details and codes can be found in the original R markdown file with the same file name.

- Select only title, journal/conference name, author, year, abstract (if any), pmid.
- Rename the column names.
- Merge and unify the columns.
- Identify if the source is web of science or pubmed.

- Duplicates check. We merged the data set by PMID in the previous step. There are 516 duplicates. Within the duplicates, we found two papers with their correction, see below. We removed papers with PMID 32817711 and 35311903.
- Unify the name of publications.

Table 4: The article with its correction identified by both of the database queries.

PMID	Source	Title
32614911	Both	Detecting rare diseases in electronic health records using machine learning and knowledge engineering: Case study of acute hepatic porphyria
32817711	Both	Correction: Detecting rare diseases in electronic health records using machine learning and knowledge engineering: Case study of acute hepatic porphyria
34505903	Both	Characterizing phenotypic abnormalities associated with high-risk individuals developing lung cancer using electronic health records from the All of Us researcher workbench
35311903	Both	Correction to: Characterizing phenotypic abnormalities associated with high-risk individuals developing lung cancer using electronic health records from the All of Us researcher workbench

4. Results after merging

After removing 28 AMIA articles, 516 duplicated articles, as well as 2 corrected article, the number of articles from each source, i.e. PubMed, Web of Science, or both are summarized in the table below.

Table 5: Number of articles extracted by each database

Source	n
Both	510
PubMed	205
WoS	135
Total	850

Analysis

Compare articles identified by Web of Science and PubMed

Figure 1 summarized the number of articles before merging from the two database, PubMed and Web of Sciences. Publications increased over years. Web of Science generally identified more articles than PubMed for JAMIA and JBI articles, while PubMed identified more PloS One and AMIA articles (Details can be found in Appendix where we summarized the number of articles across years).

This indicated that both databases can add articles that the other did not capture. It also demonstrated why we choose articles from the two databases, instead of using one.

Figure 2 and 3 summarized the number of articles across journals after merging, with the color bars indicating the articles are captured by both databases (purple), PubMed alone (blue), or Web of Science (green). From Figure 2, most of the articles were identified from both databases, with Web of Science generally captured more articles in JAMIA and JBI while PubMed captured more for PloS One.

Figure 3 indicated that most of the articles are identified from both databases while PubMed identified more than Web of Science through years.

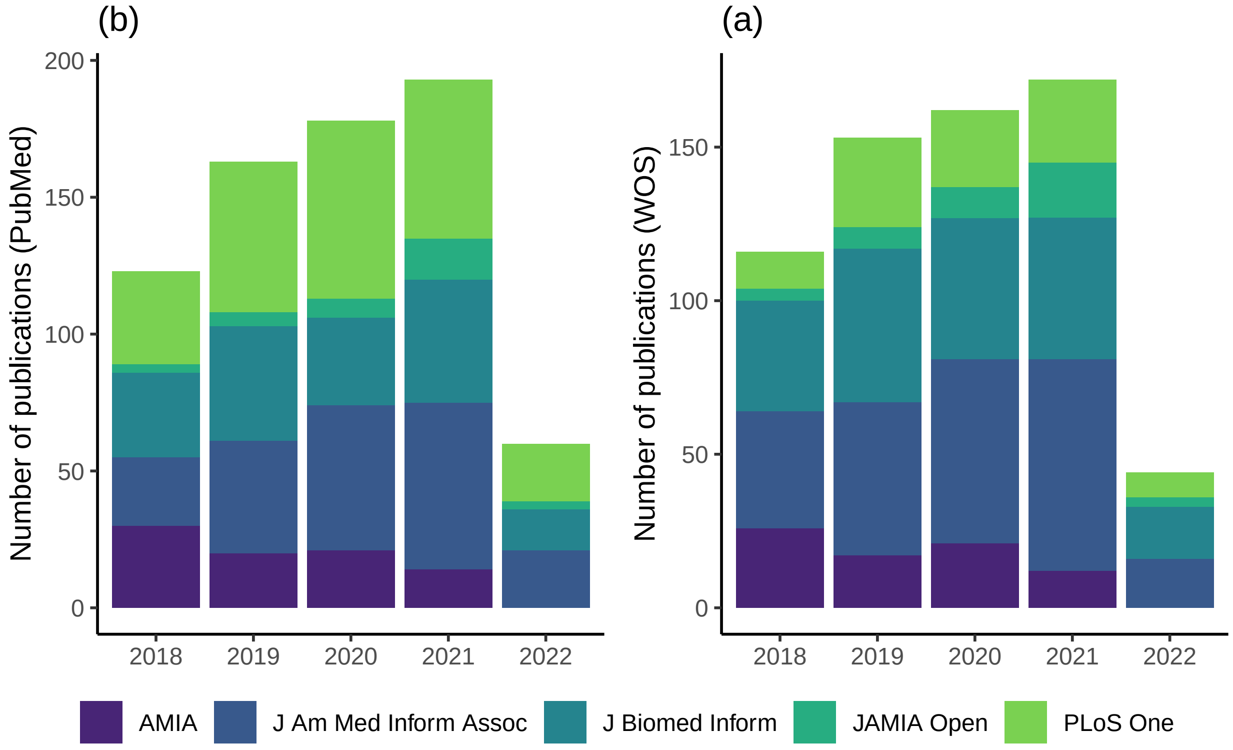

Figure 1: Number of articles across journals and years before merging. (a) number of articles extracted from PubMed. (b) Number of articles extracted from Web of Science.

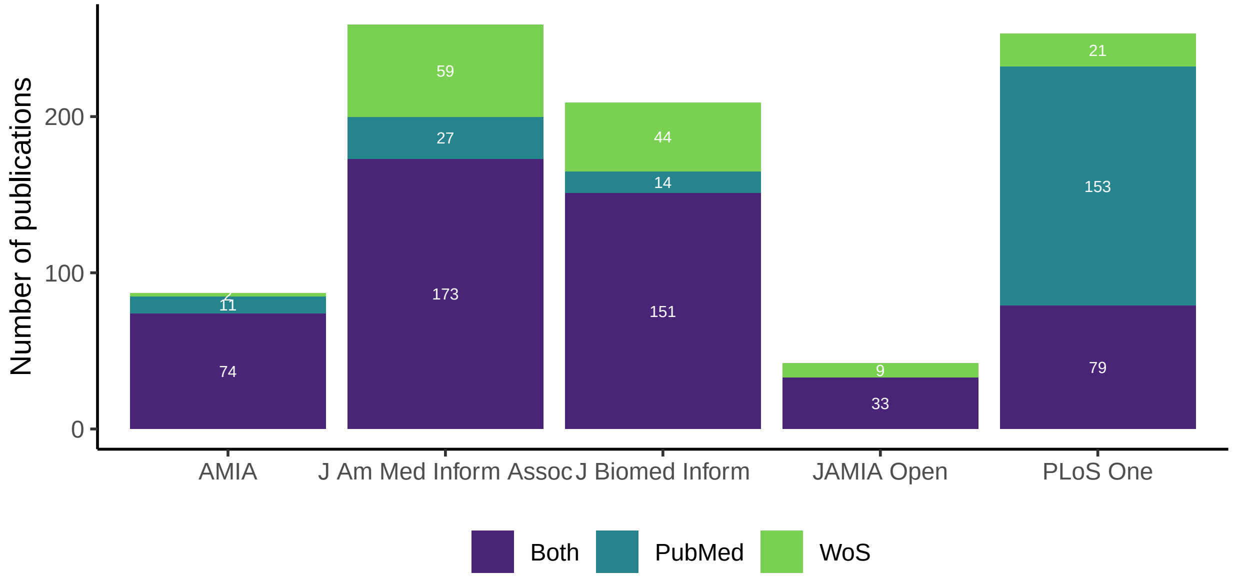

Figure 2: Number of articles across journals after merging.

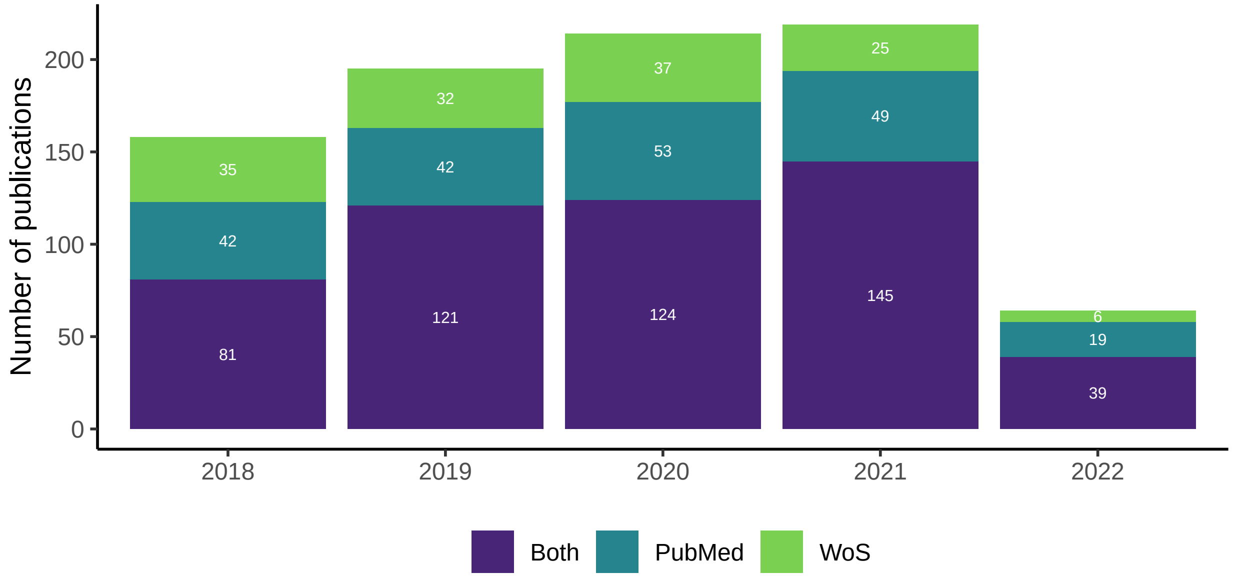

Figure 3: Number of articles across years after merging.

Compare articles across years

Figure 4 showed that the number of articles across years after merging. The number of publications increases over years.

JAMIA and JAMIA Open articles, together with a total number of 301 articles published during the four years, are identified the most. PloS One published the second most articles, with a total of 253 articles. JBI published slightly less. AMIA captured fewest articles and the number of publications is not monotonically increasing, this might suggest that not all relevant articles from the two sources are well-indexed by PubMed and Web of Science.

Summary

- 510 (60.0%) articles are captured by both of the databases, PubMed captured additional 205 articles (24.1%) and WoS captured additional 135 (15.9%), shown in Table 5. We also benefit from using the two different queries as they both captured additional articles.
- PubMed generally identified more PLoS One articles than Web of Science.
- Most articles are from JAMIA and PloS One.

We saved the list of articles after merging in the csv format, with the name 'merged_20220414.csv'. The file will be used for manually screening to include or exclude articles that are not relevant to the purpose of our scoping review. Please find more details in other R markdown scripts.

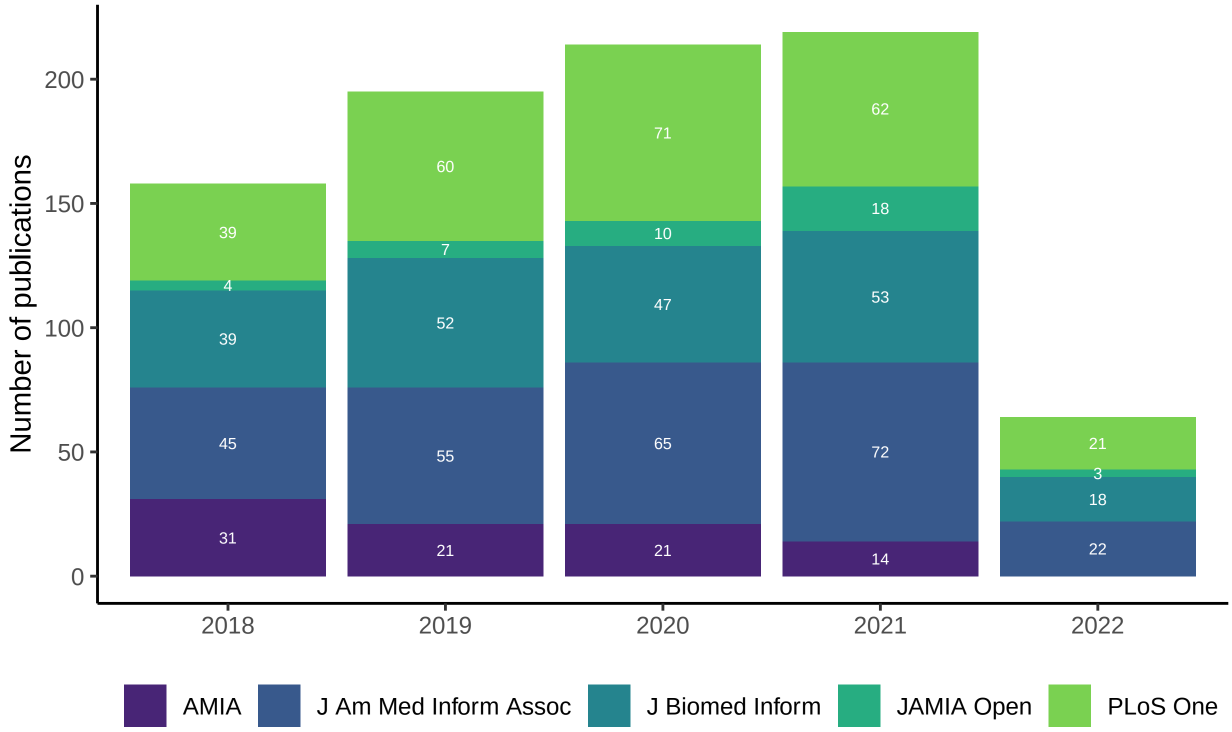

Figure 4: Number of articles across years after merging, stratified by the journal or conference

Appendix

Table 6: Number of articles stratified by journals before merging.

JournalorConference	wos	pubmed
AMIA	76	85
J Am Med Inform Assoc	233	201
J Biomed Inform	195	165
JAMIA Open	42	33
PLoS One	101	233
Total	647	717

Table 7: Number of articles stratified by years before merging.

Year	wos	pubmed
2018	116	123
2019	153	163
2020	162	178
2021	172	193
2022	44	60
Total	647	717

Table 8: Number of articles by journal/conference after merging.

JournalorConference	n
AMIA	87
J Am Med Inform Assoc	259
J Biomed Inform	209
JAMIA Open	42
PLoS One	253
Total	850

Table 9: Number of articles across years after merging.

Year	n
2018	158
2019	195
2020	214
2021	219
2022	64
Total	850