Article merging

Siyue Yang, Jessica Gronsbell

Last Updated: 10/10/2022

We extracted articles from PubMed and Web of Science published in 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)

The data of our most recent query was April 14, 2022.

Table 1: Number of articles extracted by initial query

Source	n
PubMed	745
Web of Science	651
Total	1396

Merging procedure

We next detail the key steps of merging the articles from the two databases.

1. Resolve publication time conflict in AMIA articles extracted by PubMed

We noticed that there were a large number of AMIA articles extracted by PubMed in 2018 relative to the other years.

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 publication date of AMIA articles. PubMed often extracts AMIA articles using the acceptance

date and records the acceptance data as the publication date. We checked the AMIA PubMed journal list and validated our assumptions. This time conflict does not occur in Web of Science.

In order to merge the results from two databases, we manually searched the titles in Google Scholar and corrected the year of publication for these AMIA articles (in `amia20220414.csv`). We then removed all AMIA articles accepted in 2017. Note that due to the date of our query we did not capture AMIA articles from 2022.

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. Correct errors in record information in Web of Science

The article "Extraction of Active Medications and Adherence Using Natural Language Processing for Glaucoma Patients" published in AMIA is recorded as being published in "OHSU Digital Commons". We manually changed it to "AMIA" in the next section when merging.

Two articles extracted by Web of Science did not have a PMID, "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.

We also removed duplicate articles extracted by Web of Science.

3. Merge articles from the two databases

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 (note: not extracted by PubMed), and PMID.
- Rename the column names.
- Unify the name of publications.
- Merge and unify the columns.
- Identify if the source is web of science or pubmed.
- Duplicates check. Within the duplicates, we found two papers with their correction, see below. We removed papers with PMID 32817711 and 35311903.

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. Merging results

We summarize the merged data containing the extracted articles from the two databases after removing AMIA articles with the incorrect publication date, duplicated articles, and article corrections described in the previous section.

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

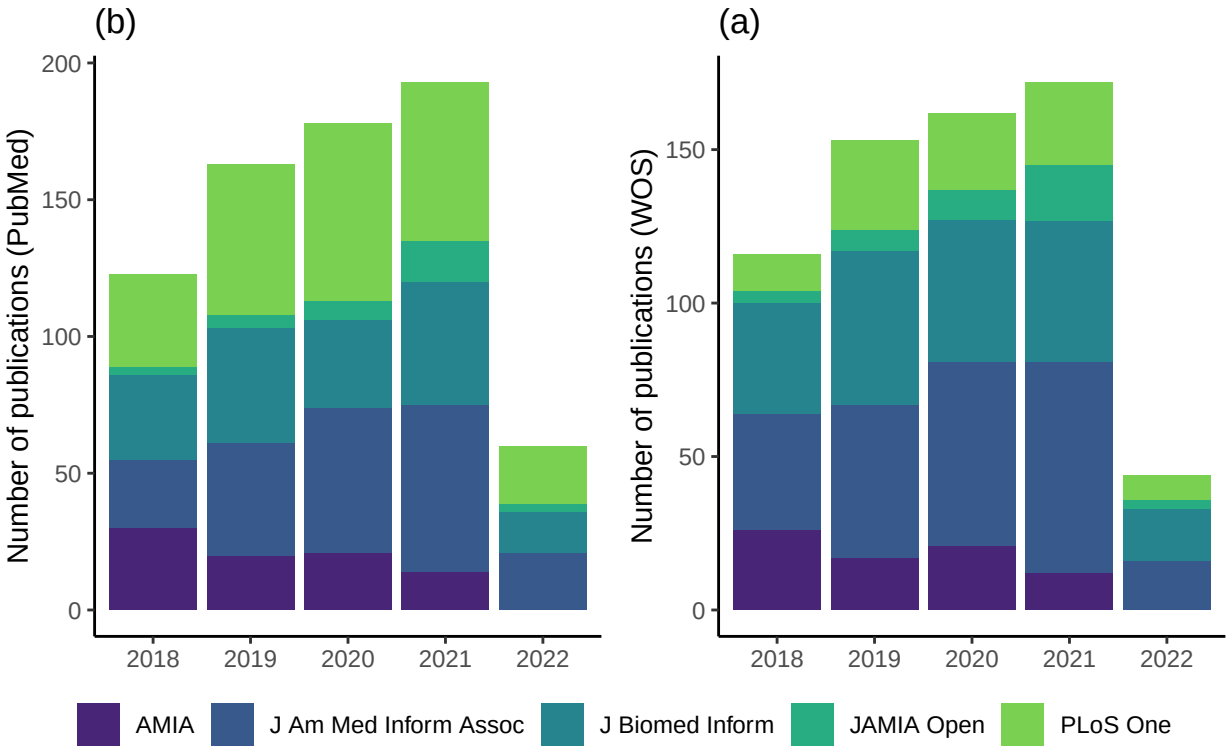

Figure 1 indicates publications increased over years. Web of Science generally identified more articles than PubMed for JAMIA and JBI articles, while PubMed identified more PloS One articles. This indicates distinct articles are captured by both databases and explains why we queried both databases.

Figures 2 and 3 summarize the number of articles across journals after merging, stratified by journal and year, respectively. From Figure 2, we see most of the articles were identified from both databases, with Web of Science generally capturing more articles in JAMIA and JBI while PubMed capturing more for PloS One. Figure 3 indicates that most of the articles are identified from both databases while PubMed identified more than Web of Science over the years.

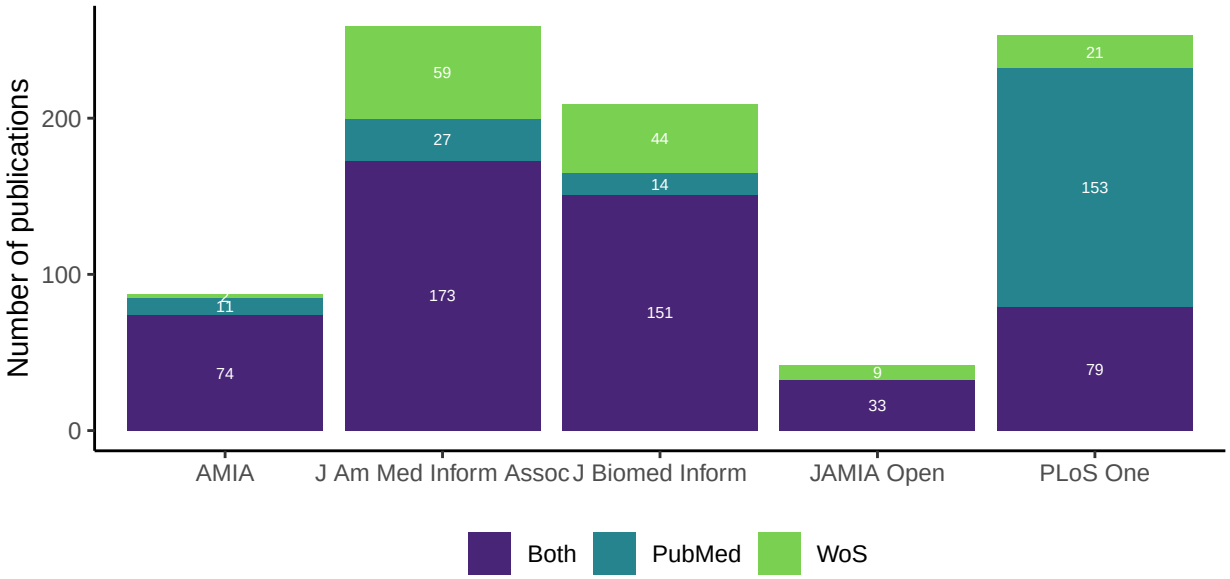

Figure 1: Number of articles stratified by journal across journals after merging.

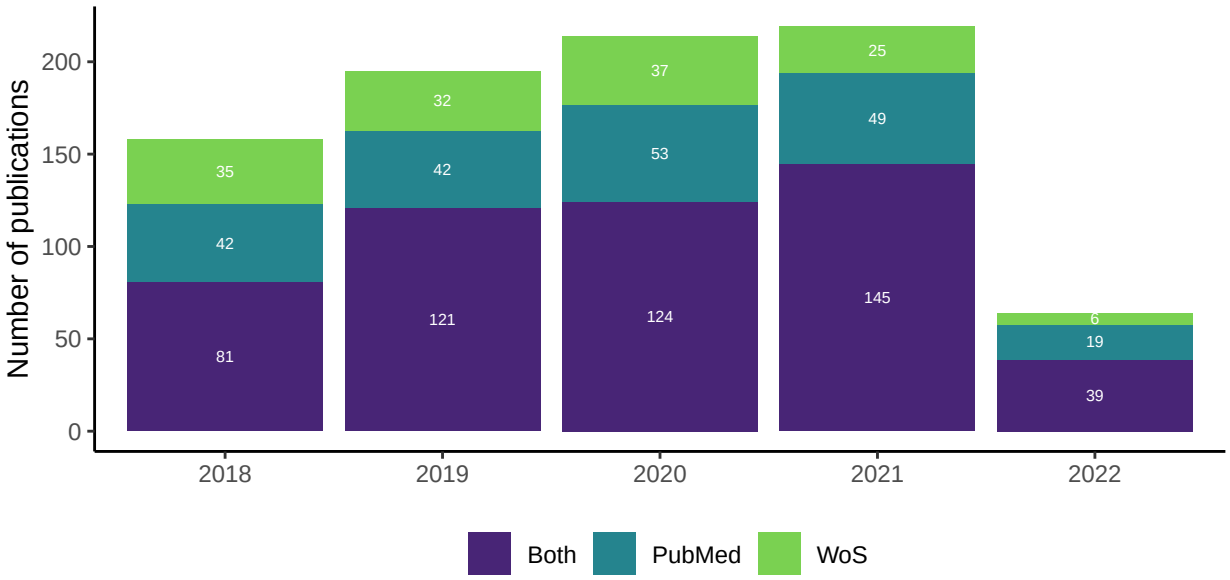

Figure 2: Number of articles stratified by year across years after merging.

Compare articles across years

Figure 4 shows 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.

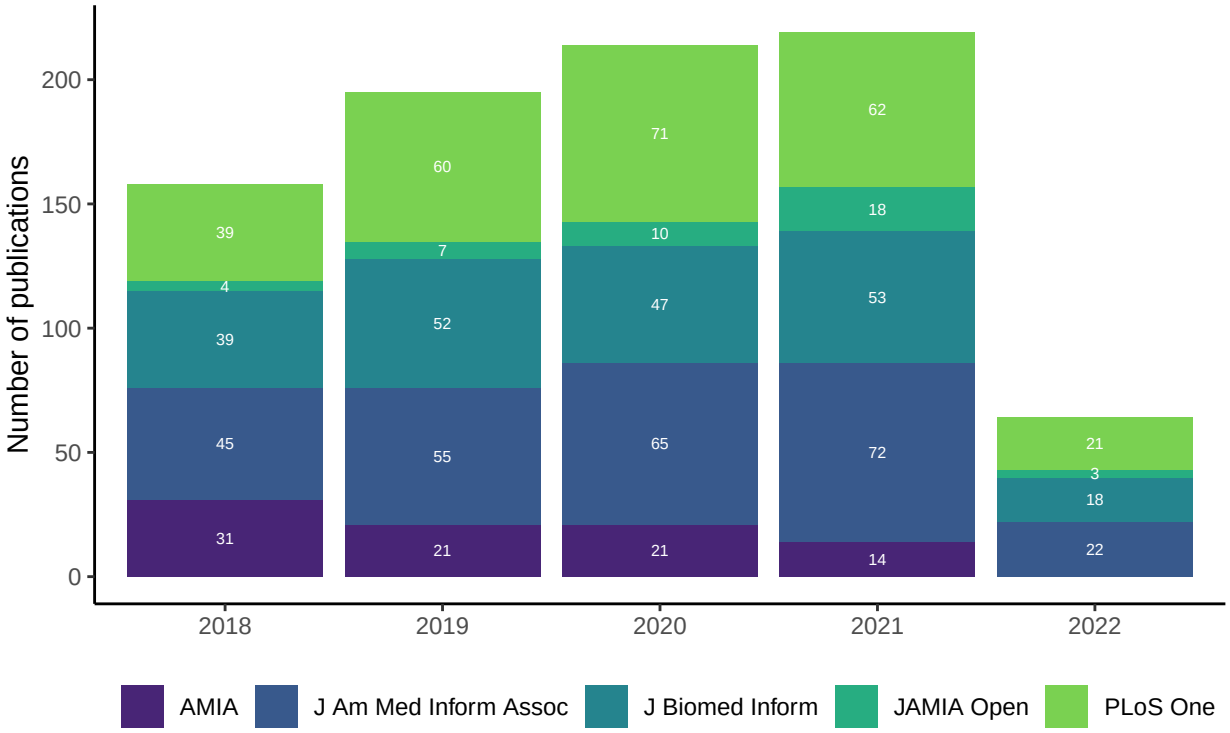

Figure 3: Number of articles across years after merging, stratified by the journal or conference

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%) (see 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'. This file was used for manual screening of articles to identify those that are relevant for our review.

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