High precision but variable recall – comparing the performance of five deduplication tools

Heidrun Janka and Maria-Inti Metzendorf
Institute of General Practice (ifam), Medical Faculty of the Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany

Abstract
Deduplication methods for multiple database searches conducted for evidence syntheses differ in terms of time invested, accuracy, and comprehensiveness of identified duplicates. Deduplication tools can significantly contribute to a more efficient conduct of the search task in evidence syntheses. Widely used tools for deduplication include reference management software (e.g. EndNote), built-in deduplication features in systematic review software (e.g. Covidence, Rayyan), and automated deduplication tools (e.g. Deduklick, SRA Deduplicator). Newer tools leverage machine learning algorithms crafted by information specialists, that encompass natural language normalization and rule-based approaches. We investigated five frequently used automated and semi-automated deduplication tools regarding their performance, core features and time efficiency in comparison to manual deduplication in EndNote using six datasets.

Key words: *systematic reviews as topic; information storage and retrieval; bibliographic databases; algorithms; software.*

Introduction

The removal of duplicate references from extensive systematic searches in different literature databases is a time-consuming and laborious process for authors or librarians supporting evidence syntheses (1). Different deduplication approaches are practiced by author teams, e.g. manual, semi-automated or automated using specialized software. These approaches vary in time-to-be-invested, completeness and accuracy of identified duplicates. Commonly used tools for a multi-step detection of duplicates are reference management programmes (e.g., EndNote (2)) and built-in deduplication features of systematic review software (e.g. Covidence (3), Rayyan (3, 4). However, deduplication processes are not made transparent in all tools and are sometimes error-prone. Newer deduplication tools such as Deduklick (5) and the SRA Deduplicator (6) use machine learning algorithms including natural language normalisation and sets-of-rules created by information specialists. Automated deduplication tools differ in the extent of the automated processes they support and in the additional manual processes required for an accurate and comprehensive detection of duplicates. While Deduklick and Covidence can be classified as "automated tools" in the sense that no additional manual control is necessary for the deduplication process (except file preparation, e.g. creating RIS-files preceding the upload process) – in tools like SRA Deduplicator and Legacy Rayyan an additional manual control of system-detected duplicates is essential, therefore considered semi-automated tools. We aimed to compare and evaluate the core features, performance, transparency and time efficiency of five frequently used manual, semi-automated and automated deduplication tools: EndNote, Covidence, Legacy Rayyan, Deduklick and SRA Deduplicator.

Methods

We used six different datasets by compiling database search results from six Systematic Reviews (*Table 1*) covering various health topics and varying in size between 300 to 1000 records. The records had previously been retrieved from bibliographic databases (MEDLINE via PubMed or Ovid, CENTRAL, CINAHL, LILACS,

Address for correspondence: Heidrun Janka, Institute of General Practice (ifam), Medical Faculty of the Heinrich-Heine-University Düsseldorf, Building 17.11.02, Moorenstr. 5, 40225 Düsseldorf, Germany.
E-Mail: heidrun.janka@med.uni-duesseldorf.de

Journal of EAHIL 2024; Vol. 20 (1): 12-17. doi: 10.32384/jeahil20607

Systematic Review Topics	Databases searched	References retrieved
Fitness to drive in dementia	MEDLINE, CINAHL, CENTRAL, PsycInfo	414
Interventions for people with type 2 diabetes mellitus fasting during Ramadan	MEDLINE, CENTRAL, PsycINFO, CINAHL, ClinicalTrials.gov, WHO ICTRP Database	375
JAK inhibitors for the treatment of COVID-19 patients	CCSR, Web of Science, WHO COVID-19, US Dep. VA	344
Glucagon-like peptide (GLP)-1 analogues as add-on to insulin for adults with type 1 diabetes mellitus	MEDLINE, CENTRAL, ClinicalTrials.gov, WHO ICTRP Database	833
Vegan diet for overweight or obese adults	MEDLINE, CENTRAL, LILACS, Web of Science, ClinicalTrials.gov, WHO ICTRP Database	1002
Vitamin D supplementation for obese adults undergoing bariatric surgery	MEDLINE, CENTRAL, LILACS, ClinicalTrials.gov, WHO ICTRP	966

Table 1. *Systematic Review topics.*

PsycInfo, Web of Science, Cochrane Covid-19 Study Register and also from trials registers (ClinicalTrials.gov, WHO ICTRP Database). We tested each dataset on Deduklick, SRA Deduplicator ("focused" and "relaxed" algorithm), Covidence and Legacy Rayyan to compare the deduplication performance against the manual procedure in EndNote. The manual deduplication used as reference standard was conducted by an information specialist using a 12-step algorithm (7). It was defined as obtaining the same results twice after undertaking two independent deduplication procedures. The core features investigated for each tool included data processing (upload process, data formats accepted, the delivery of deduplication reports informing on all bibliographic details from the datasets removed resp. retained, as well as the database origins displayed in a flow diagram, and on separate export files containing duplicates as well as the deduplicated results), transparency of the deduplication process (e.g. transparency about the database fields being compared as well as the display of all available metadata for identified duplicates) and additional options like the possibility to define keeping bibliographic records from preferred databases. The time-investment required for the deduplication process was measured in minutes and comprised the time for the file upload, the system-detected deduplication and the additional manual deduplication required.

Results

Comparison of the deduplication performance

Table 2 presents the average scores of system-detected duplicates from six datasets for all tools: automated (Deduklick, Covidence) and semi-automated tools (SRA Deduplicator, Legacy Rayyan) in comparison to manual deduplication. For definitions of precision and recall of a tool's deduplication performance see *Figure 1*.

While on average precision for identifying duplicates was very high in all tools except for Rayyan, the recall (sensitivity) varied substantially. Deduklick and Legacy Rayyan were the most sensitive tools according to our

Deduplication tool	Precision ⌀	Recall ⌀
Covidence	100%	76.8%
Deduklick	100%	96.2%
SRA Deduplicator (focused)	99.8%	86.9%
SRA Deduplicator (relaxed)	100%	73.9%
Rayyan	95.5%	99.1%

Table 2. *Average scores for deduplication performance.*

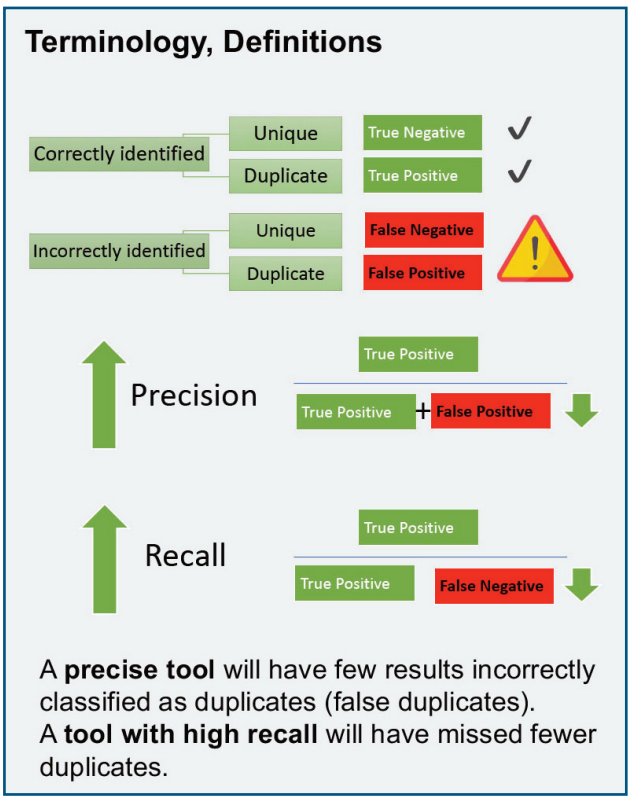

Fig. 1. *Precision and recall for the performance of deduplication tools.*

tests. However, Legacy Rayyan also detected false positive references (therefore precision was lower). The lowest recall (i.e. highest rate of missed duplicates) was measured for SRA Deduplicator ("relaxed"), followed by Covidence. As the SRA Deduplicator is considered a semi-automated tool, additional manual control is generally recommended to ensure better recall. In Covidence, low recall might in part be explained by the limited numbers of database fields being compared for duplicate detection – in contrast to Deduklick in which ten database fields are used. For further details on the comparison of the tools' core features see *Table 3*.
Looking at the deduplication performance in single datasets, low recall was observed more frequently in datasets with larger proportions of records from trials registers (probably because of metadata quality) in comparison to data originating from bibliographic databases.

Comparison of core features

Data processing, transparency of the deduplication process and additional features offered by the tools are presented in *Table 3*. Concerning the data formats accepted for import, EndNote and Legacy Rayyan seem to be the most flexible tools followed by SRA

Features	Covidence	Deduklick	SRA Deduplicator	Legacy Rayyan	EndNote
Deduplication method	Automated	Automated	Semi-Automated	Semi-Automated	Manual
Manual checks	Additional manual check possible	–	Additional manual check recommended	Manual check necessary	–
Data formats accepted	RIS, EndNote XML, PubMed nbib	RIS - preprocessed in EndNote	RIS, EndNote XML, PubMed nbib	RIS, EndNote enw, BibTeX, CSV, PubMed XML, PubMed nbib, CIW	All bibliographic formats
Deduplicated + duplicates files for download	x	✔	✔	✔	✔
Duplicates Report	✔	✔	x	x	x
Database preference + ranking (for import)	x	✔	x	x	(✔)
Database fields checked	TI, AU, YEAR, VOL	TI, AU, TA, DOI, YEAR, ISSN, VOL, PG, URL, AN	Focused algorithm: 10 fields (not named); Relaxed algorithm: 5 fields (not named)	TI, AU, TA, YEAR	12-step algorithm with different field combinations

Table 3. *Comparison of the core features of deduplication tools*

Deduplicator and Covidence. Deduklick currently only accepts merged RIS-files that need to be preprocessed in reference management software. The download of deduplicated reference files and the files containing the duplicates is possible in all tools except Covidence. The latter displays a list of potential duplicate references which can be manually checked, but only with limited bibliographic information (e.g. database sources are missing in the record). The duplicate lists can neither be saved nor downloaded. Comprehensive bibliographic database records for duplicates can be downloaded from Deduklick, SRA Deduplicator and EndNote. Detailed deduplication reports are only available from Deduklick, in addition to flow diagrams displaying the number of references per source before and after deduplication. An additional advantage of this tool is the possibility to customize the database ranking: Deduklick retains unique records from databases providing the most complete bibliographic data and removes duplicates from other databases / sources. The database ranking has been determined by information specialists of the University of Bern, but Deduklick offers customizing this list upon request.

Comparison of time efficiency

Deduklick and Covidence are the fastest tools for deduplication, including file upload and an automated detection of duplicates within 2-5 minutes, depending on the size of the uploaded files (*Figure 2*). All other tools need more time because additional manual work is required. The SRA Deduplicator offers a "relaxed" algorithm which is designed for people who want to spend minimal time with checking the results manually, according to the producers. The risk of mislabeling non-duplicates is low, however, at the expense of missing a small number of duplicates. Legacy Rayyan proved to be the most time-consuming tool, requiring more than one hour of additional manual work, depending on the file size, due to its very sensitive similarity score. It is noteworthy that using this tool required even more time than carrying out manual deduplication with EndNote. In summary, *Table 4* provides an overview of the best-performing deduplication tools according to the different criteria investigated in this study.

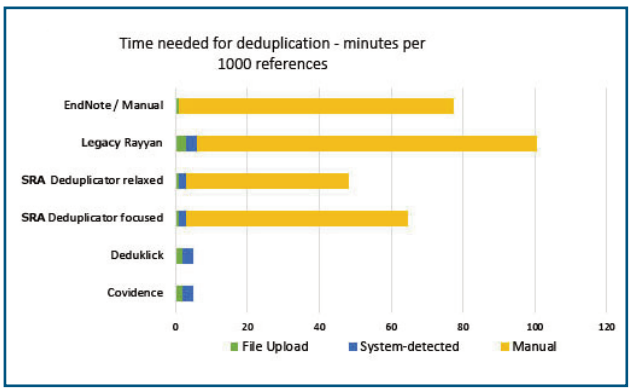

Fig. 2. *Average time efficiency of deduplication tools accross 6 datasets.*

Criteria	Best performing tools
Precision	Deduklick, SRA Deduplicator (focused + relaxed), Covidence
Recall	Legacy Rayyan, Deduklick, SRA Deduplicator (focused)
Time Efficiency	Deduklick, Covidence
Core Features	Deduklick, SRA Deduplicator (focused + relaxed)

Table 4. *Summary of all categories investigated.*

Discussion

In our tests, Deduklick, the SRA Deduplicator ("focused" and "relaxed" algorithm) and Covidence could be identified as the most precise tools for duplicate detection, whereas highest recall was achieved by Deduklick and Legacy Rayyan – in comparison to manual deduplication in EndNote (reference standard). An earlier investigation conducted by McKeown and Mir (8), found Covidence and Ovid to be the most accurate tools (96% and 97%, respectively) for duplicate detection, with Covidence and Ovid* possessing the highest specifity (100%), while Legacy Rayyan demonstrated the highest sensitivity (96%) (8).

The pros and cons of using reference management software for deduplication were also investigated by McKeown and Mir (8). They evaluated EndNote X9,

*As a host and provider of bibliographic databases such as MEDLINE, Embase, PsycInfo etc., Ovid offers a built-in deduplication function on its platform, which enables the user to detect duplicates from searches across various databases available via Ovid. The advantage of deduplication in Ovid is that metadata from hosted databases are structured in a similar way, making it easier to identify duplicates, the disadvantage being that deduplication is only possible for databases available on Ovid.

Mendeley and Zotero. In each tool, the system´s default settings for deduplication were used, but no additional manual deduplication algorithms applied, e.g. Bramer (2) or Wright (7), therefore the results are different and not comparable to ours. Differences in the accuracy of deduplication in their investigation compared to our study may also be explained by the different composition of databases used in the datasets, most importantly, the omission of data from trials registers, which were included in our datasets.

The SRA Deduplicator is one of several tools integrated in the SR Accelerator tool which was designed at Bond University, Australia. The tool aims to speed up several of the processes of systematic review production while maintaining a high degree of accuracy (6). The SRA deduplicator is freely available as part of the suit of tools and offers two different deduplication algorithms ("focused" and "relaxed"). While additional manual deduplication is generally recommended for this semi-automated tool, the "relaxed" algorithm can also be used with its default setting, at the risk of missing a few duplicates but without false positive records labeled (6). In our investigation, the average recall was 74% when applying the "relaxed" algorithm, which results in an average of 26% of records of manual work.

Concerning time efficiency, the two automated tools, Deduklick and Covidence, demonstrate the fastest performance – followed by the SRA Deduplicator. According to Forbes *et al.* (6), the time needed for deduplication of 10 different datasets taken from Cochrane Reviews, with reference numbers ranging between 813 and 3912, the SRA Deduplicator was on the average 330% faster compared to the manual deduplication method in EndNote. This contrasts with our findings, as the average time savings measured by us were around 75% with the SRA Deduplicator (only using system-detected duplicates) compared to EndNote. However, after conducting the additional manual deduplication, the time savings resulted in only 15-20% when compared to manual deduplication in EndNote.

Conclusions

We investigated five frequently used automated and semi-automated deduplication tools regarding their performance, core features and time efficiency in comparison to manual deduplication in EndNote as reference standard. Six datasets, derived from Systematic Reviews and composed of heterogenous bibliographic data from medical databases and trials registers, were tested on all tools. We observed high precision (95-100%) in detecting duplicates for all tools, but variable recall (74-99%). Time efficiency varied substantially between two to five minutes (Deduklick, Covidence) and more than one hour (Legacy Rayyan), depending on the size of files deduplicated, the proportion of automated processes versus remaining manual work, and on the metadata composition of the datasets investigated. Core features that differ between the tools are data formats accepted, the possibility of downloading duplicates and deduplicated files as well as the availability of deduplication reports.

Note: Since our tests in spring 2023 and the writing of this article in February 2024, a new version of Rayyan has been released whose deduplication features have improved. In our text we refer to the "Rayyan Legacy" version.

Conflicts of interest: HJ was involved in the development and testing of the Deduklick tool.

Acknowledgements

The authors would like to thank Juan VA Franco and Brenda Bongaerts for their thoughtful comments on this project, which was first presented at the Cochrane Colloquium 2023 in London.

Submitted on invitation.
Accepted on 28 February 2024.

.

REFERENCES

1. Rathbone J, Carter M, Hoffmann T, Glasziou P. Better duplicate detection for systematic reviewers: evaluation of Systematic Review Assistant-Deduplication Module. Syst Rev. 2015 Jan 14;4(1):6. doi: 10.1186/2046-4053-4-6. PMID: 25588387.
2. Bramer WM, Milic J, Mast F. Reviewing retrieved references for inclusion in systematic reviews using EndNote. J Med Libr Assoc. 2017;105(1):84-7.

3. Kellermeyer L, Harnke B, Knight S. Covidence and Rayyan. J Med Libr Assoc. 2018 Oct;106(4):580-3. doi: 10.5195/jmla.2018.513. Epub 2018 Oct 1. PMCID: PMC6148615.
4. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev. 2016 Dec 5;5(1):210. doi: 10.1186/s13643-016-0384-4. PMID: 27919275; PMCID: PMC5139140.
5. Borissov N, Haas Q, Minder B, Kopp-Heim D, von Gernler M, Janka H, et al. Reducing systematic review burden using Deduklick: a novel, automated, reliable, and explainable deduplication algorithm to foster medical research. Syst Rev. 2022;11(1):172.
6. Forbes C, Clark J, Greenwood H. Automation of Duplicate Detection for Systematic Reviews. ICML 2022. Available online: https://icml2022.org/assets/documents/icml2022%20proceedings/D14-ICML%20AHILA2022-C%20FORBES-FINAL%20PAPER.pdf
7. Wright J. University of Leeds Checking for Duplicates Guidance. 2016. Available online: https://information-specialists.leeds.ac.uk/wp-content/uploads/sites/71/2019/03/Duplicate_checking_guidance.pdf
8. McKeown S, Mir ZM. Considerations for conducting systematic reviews: evaluating the performance of different methods for de-duplicating references. Syst Rev. 2021;10(1):38.