Automation of Duplicate Detection for Systematic Reviews

Connor Forbes
Institute for Evidence-Based Healthcare, Bond University, Gold Coast, Australia.
E-mail address: cforbes@bond.edu.au

Justin Clark
Institute for Evidence-Based Healthcare, Bond University, Gold Coast, Australia.
E-mail address: jclark@bond.edu.au

Hannah Greenwood
Institute for Evidence-Based Healthcare, Bond University, Gold Coast, Australia.
E-mail address: hgreenwo@bond.edu.au

Abstract:

In this paper, we investigate the use of an automation tool, the "Deduplicator" for removing duplicate articles from a multi-database search for systematic reviews. We compare the Deduplicator to a manual method using EndNote to deduplicate articles by testing the performance on 10 previous Cochrane systematic reviews. Two researchers each performed deduplication on the ten libraries. For five of those libraries one researcher used the Deduplicator, while the other one performed manual deduplication with EndNote. They then switched methods for the remaining five libraries.

With the Deduplicator tool, the average time to deduplicate a library was 8.2 minutes compared to 27 minutes with the manual method. Researchers averaged 299.53 references per minute when using Deduplicator compared to 99.22 references/minute with the manual method. Deduplicator achieved an average accuracy of 99.82% compared to 99.70% for the manual method. This demonstrates evidence that using the Deduplicator for duplicate article detection reduces the time taken to deduplicate, while maintaining or improving accuracy compared to using EndNote.

Keywords: Deduplication, Systematic Reviews, Duplicate Articles, Searching, Automatic.

Introduction:
Systematic Reviews are considered the best way to answer a research question. However, they are resource intensive; taking on average, five staff, 67 weeks to complete (Borah, Brown, Capers and Kaiser, 2017) at an average cost of USD $141,000 (Michelson and Reuter, 2019). To overcome this resource burden, Systematic Review Automation (SRA) tools have been developed to improve the speed of Systematic Review (SR) tasks, without compromising quality (Beller et al., 2018). A time-consuming task is to remove duplicate records from search results. This can take even experienced searchers hours to complete. We have designed an SRA tool the "Deduplicator" with the goal of speeding up this process, while also maintaining a high degree of accuracy. This paper aims to evaluate the effectiveness of the Deduplicator tool at minimising time spent screening while maximising accuracy. The Deduplicator tool is freely accessible online at the following link (https://sr-accelerator.com/#/deduplicator).

Background:
When performing a systematic review, it has become standard to perform a multi-database search when finding evidence to ensure evidence is unbiased and complete (McKibbon, Wilczynski, Eady and Marks, 2009). However, databases may list the same reference twice, meaning that the citation appears more than once, also known as a duplicate article (Kwon, Lemieux, McTavish and Wathen, 2015). This complete list of references is known as a "library" and before performing screening in a systematic review, it is ideal to remove any duplicate articles from the library to minimise redundant time spent screening same article twice. This process is referred to as deduplication.

Methods:
Ten libraries were randomly chosen from past Cochrane systematic reviews, published in the last five years (Jan 2017 – Sep 2021) where there were at least two databases listed in the review. Searches were run as they were written in the review, with no date limits applied in all listed searches. Searches were run in all bibliographic databases listed in the review but were not performed in databases not available at Bond University. Specialised registers, trial registries and grey-lit databases were also excluded. If searches returned less than 500 references or greater than 10,000 references, the review was discarded.

To evaluate the Deduplicator we will compare deduplication done manually and done with the Deduplicator on the following outcomes: 1) time required to deduplicate; 2) numbers of duplicates missed; 3) number of non-duplicates removed. Two screeners (HG & JC) will independently deduplicate 10 sets of search results. The first screener will do sets one to five manually, then sets six to 10 with the Deduplicator. The second screener will do the opposite, e.g., sets one to five with the Deduplicator, then sets six to 10 manually (see Table 1). Here, manual deduplication is defined as using Endnote on an adapted deduplication method originally proposed by Bramer (Bramer et al., 2016). In the Deduplicator, the "Balanced" algorithm was selected for evaluation. Researchers also timed themselves on how long it took to perform deduplication on each library.

Systematic Review (Author Year)	Hannah Greenwood (Method)	Justin Clark (Method)
Lorentzen 2020	Manual	Deduplicator
Alebed 2020	Deduplicator	Manual
Dawson 2021	Manual	Deduplicator

Wiffen 2017	Deduplicator	Manual
Kamath 2020	Manual	Deduplicator
Ghobara 2017	Deduplicator	Manual
Bennett 2018	Manual	Deduplicator
Hannon 2021	Deduplicator	Manual
Roberts 2020	Manual	Deduplicator
Jaschinski 2018	Deduplicator	Manual

Table 1: Distribution of manual vs Deduplicator methods between researchers

For resolving the errors, both participants libraries were compared against each other. Any discrepancies between the duplicate results for both libraries were manually checked and verified by consensus between two authors (HG & CF). If a reference was incorrectly classified as a duplicate when it is in fact a unique article, it is labelled as a "false positive", while a duplicate which was missed is marked as a "false negative".

Results:

While testing on the libraries, Deduplicator was on average 330% faster compared to the manual EndNote method (8.2 minutes vs 27.0 minutes respectively). The median time of the Deduplicator was 6.5 minutes compared to 25 minutes for the manual method. The results for each library are displayed in Table 2. Deduplicator averaged 299.52 references per minute while the manual method averaged 99.22 references per minute (see Table 3).

Systematic Review (Author Year)	Study Size (Number of References)	Deduplicator (Minutes)	Manual (Minutes)
Lorentzen 2020	813	4	30
Alebed 2020	1479	14	15
Dawson 2021	3912	6	76
Wiffen 2017	1028	9	7
Kamath 2020	1785	4	36
Ghobara 2017	1807	7	24
Bennett 2018	2111	4	35
Hannon 2021	1061	5	6
Roberts 2020	3181	9	15
Jaschinski 2018	2447	20	26
Average	**1962.4**	**8.2**	**27.0**
Median	**1796**	**6.5**	**25.0**

Table 2: Full break-down of the time taken to deduplicate with each method

Author	Deduplicator (references/minute)	Manual (references/minute)	Average
HG	162.51	80.11	121.30
JC	436.53	118.33	277.44
Average	299.53	99.22	

Table 3: Average number of references deduplicated per minute (by author)

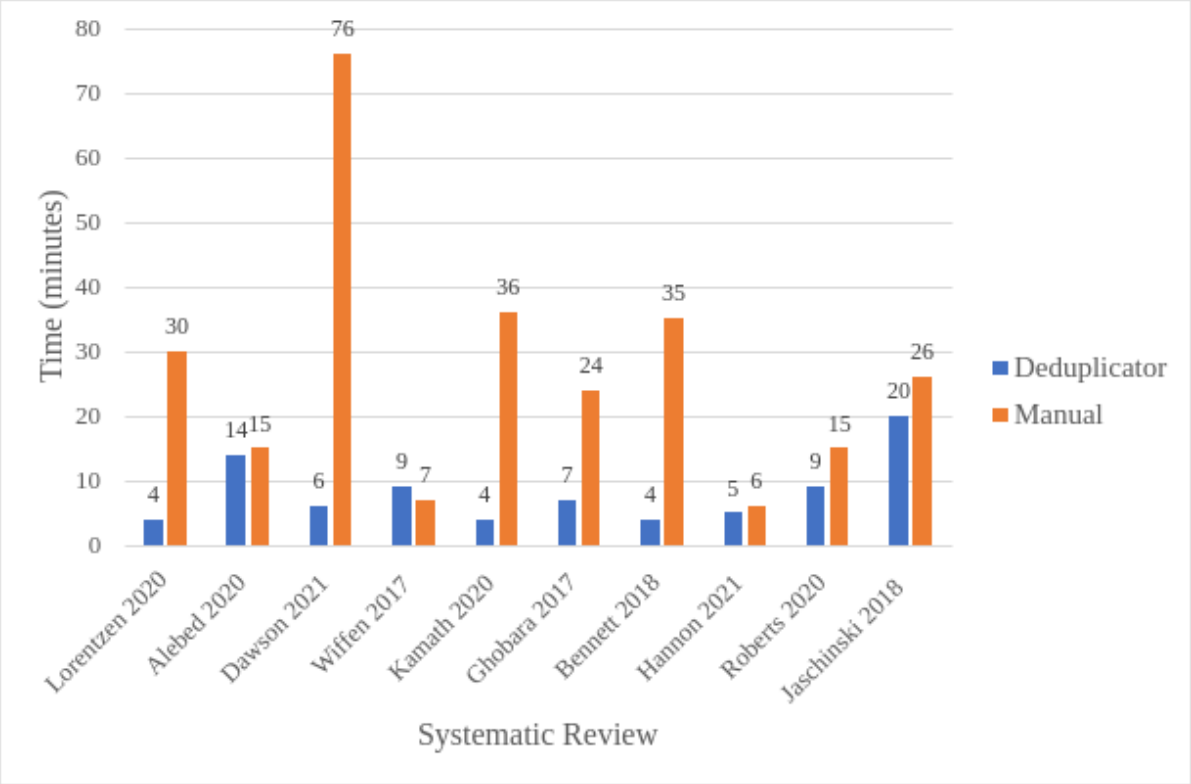

Figure 1: Time taken to deduplicate each systematic review

On average, Deduplicator made 3.3 errors per systematic review, while the manual method had an average of 6.2 errors per systematic review. The median number of errors for Deduplicator and the manual method were 3 and 5, respectively. This resulted in an average accuracy for Deduplicator of 99.82% compared to 99.70% for the manual Endnote method. These results are explored further in Table 4.

		Deduplicator			Manual		
SR	**Total References**	**False Negative**	**False Positive**	**Errors**	**False Negative**	**False Positive**	**Errors**
Lorentzen 2020	813	0	0	0	1	1	2
Alebed 2020	1479	5	1	6	3	5	8
Dawson 2021	3912	0	2	2	5	2	7
Wiffen 2017	1028	0	1	1	0	0	0
Kamath 2020	1785	2	0	2	1	1	2
Ghobara 2017	1807	4	2	6	2	3	5
Bennett 2018	2111	2	1	3	2	2	4
Hannon 2021	1061	0	3	3	3	2	5
Roberts 2020	3181	0	3	3	4	12	16
Jaschinsk	2447	5	2	7	8	5	13

i 2018							
Average	1962.4	1.8	1.5	3.3	2.9	3.3	6.2
Median	1796.0	1.0	1.5	3.0	2.5	2.0	5.0

Table 4: Comparison of number of errors for each library

Discussion:

The results show evidence that Deduplicator can perform deduplication to a high degree of accuracy, while also reducing the time needed to perform deduplication compared to using the manual EndNote method. The most important statistic is the low rate of false positives when using Deduplicator. False positives are the least desirable error as it means that an article which may contain relevant evidence to the systematic review protocol is discarded. The low false positive rate of Deduplicator (average 1.5 references per library) is desirable here as more relevant evidence improves the quality of a systematic review.

One interesting result is that for the "Wiffen, 2017" systematic review, Deduplicator was slower to deduplicate compared to the manual EndNote method. The explanation for this may be the difference in speeds between the two researchers, where JC on average deduplicated 277.44 references per minute compared to 121.30 for HG (see Table 3). Because the Wiffen library is relatively small (1028 references), it is possible that JC's extra experience allowed him to deduplicate the library quicker in EndNote compared to how quickly HG could do it in Deduplicator. This difference in deduplication speed/accuracy between authors is one limitation of this study design, and while it is partially mitigated by the equal split of methods used by each author; it is not possible to eliminate this bias entirely. Despite this, independent analysis of each author revealed that Deduplicator increased the number of references they could deduplicate per minute (see Table 3).

Furthermore, another limitation behind this study design is that if both authors incorrectly classified an article, it would not be counted as an error. However, as this is a head-to-head comparison, this limitation would not affect the comparison in accuracy between the two methods.

It should also be noted that there are multiple other deduplication tools available to perform duplicate detection. A study run by McKeown on various deduplication methods found that the highest accuracy deduplication tools were Ovid and Raayan, both achieving an accuracy of 0.97 (McKeown and Mir, 2021). While no direct comparison can be made to the existing literature due to the difference in datasets, the accuracy the Deduplicator achieved (average 0.99) may warrant a future comparison between other deduplication tools using a consistent dataset.

Conclusion:

This investigation shows evidence that the Deduplicator is quicker for the deduplication of articles compared to manual EndNote methods, without sacrificing any accuracy. Deduplicator eliminates the need to research a deduplication method in EndNote, as the tool provides a preconfigured strategy for the user. This both allows an easier point of entry for new researchers to begin deduplicating, as well as providing time-saving bonuses for more experienced researchers without any loss of accuracy.

Acknowledgments
We would like to acknowledge Dr Matt Carter for his work he has done on the coding of the Deduplicator algorithm. We would also like to thank our numerous sources of funding for development of the Deduplicator.

References
Beller, E., Clark, J., Tsafnat, G., Adams, C., Diehl, H., Lund, H., Ouzzani, M., Thayer, K., Thomas, J., Turner, T., Xia, J., Robinson, K. and Glasziou, P., 2018. Making progress with the automation of systematic reviews: principles of the International Collaboration for the Automation of Systematic Reviews (ICASR). *Systematic Reviews*, 7(1), p.77.

Borah, R., Brown, A., Capers, P. and Kaiser, K., 2017. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open*, 7(2), p.e012545.

Bramer, W., Giustini, D., de Jonge, G., Holland, L. and Bekhuis, T., 2016. De-duplication of database search results for systematic reviews in EndNote. *Journal of the Medical Library Association: JMLA*, 104(3), pp.240-243.

Kwon, Y., Lemieux, M., McTavish, J. and Wathen, N., 2015. Identifying and removing duplicate records from systematic review searches. *Journal of the Medical Library Association: JMLA*, 103(4), pp.184-188.

McKeown, S. and Mir, Z., 2021. Considerations for conducting systematic reviews: evaluating the performance of different methods for de-duplicating references. *Systematic Reviews*, 10(1).

McKibbon, A., Wilczynski, N., Eady, A. and Marks, S., 2009. *PDQ evidence-based principles and practice*. 2nd ed. Shelton, Conn.: People's Medical Pub. House, pp.192-207.

Michelson, M. and Reuter, K., 2019. The significant cost of systematic reviews and meta-analyses: A call for greater involvement of machine learning to assess the promise of clinical trials. *Contemporary Clinical Trials Communications*, 16, p.100443.