METHODOLOGY

Open Access

Automation of duplicate record detection for systematic reviews: Deduplicator

Connor Forbes[1*], Hannah Greenwood[1], Matt Carter[1] and Justin Clark[1]

Abstract

Background To describe the algorithm and investigate the efficacy of a novel systematic review automation tool "the Deduplicator" to remove duplicate records from a multi-database systematic review search.

Methods We constructed and tested the efficacy of the Deduplicator tool by using 10 previous Cochrane systematic review search results to compare the Deduplicator's 'balanced' algorithm to a semi-manual EndNote method. Two researchers each performed deduplication on the 10 libraries of search results. For five of those libraries, one researcher used the Deduplicator, while the other performed semi-manual deduplication with EndNote. They then switched methods for the remaining five libraries. In addition to this analysis, comparison between the three different Deduplicator algorithms ('balanced', 'focused' and 'relaxed') was performed on two datasets of previously deduplicated search results.

Results Before deduplication, the mean library size for the 10 systematic reviews was 1962 records. When using the Deduplicator, the mean time to deduplicate was 5 min per 1000 records compared to 15 min with EndNote. The mean error rate with Deduplicator was 1.8 errors per 1000 records in comparison to 3.1 with EndNote. Evaluation of the different Deduplicator algorithms found that the 'balanced' algorithm had the highest mean F1 score of 0.9647. The 'focused' algorithm had the highest mean accuracy of 0.9798 and the highest recall of 0.9757. The 'relaxed' algorithm had the highest mean precision of 0.9896.

Conclusions This demonstrates that using the Deduplicator for duplicate record detection reduces the time taken to deduplicate, while maintaining or improving accuracy compared to using a semi-manual EndNote method. However, further research should be performed comparing more deduplication methods to establish relative performance of the Deduplicator against other deduplication methods.

Keywords Deduplication, Systematic review, Duplicate article, Duplicate record, Searching, Automatic

Background

Systematic reviews are considered the best way to answer a research question using synthesised data; however, they can require a substantial investment of time and resources [1, 2]. On average, they take 67 weeks and cost USD $141,000 [3]. However, there are cases of systematic reviews being performed in 11 workdays by using a modified methodology that utilises systematic review automation tools [4, 5]. These systematic review automation tools have been developed with the goal of improving the speed of systematic reviews without compromising their rigour and quality [6].

One of the initial key tasks to conduct a systematic review is to find all potentially relevant studies by searching across multiple databases [7]. Due to the same journals being indexed in multiple databases, large numbers of duplicate records are frequently returned. Before the records can be assessed for relevance by reviewers (a

*Correspondence:
Connor Forbes
cforbes@bond.edu.au
[1] Institute for Evidence-Based Healthcare, Bond University, Gold Coast, Australia

process called screening), the duplicate records must be removed. This process is referred to as deduplication.

There are multiple methods to deduplicate records retrieved from searching for systematic reviews. One method of deduplication utilised by researchers is to use a semi-manual method, combining software such as EndNote with human checking, although this method is still prone to errors [8]. Despite deduplication being a routine task in systematic reviews, there is little consensus about the best method of deduplication [8]. Although there have been attempts to standardise semi-manual deduplication methods, they rely on the steps being applied consistently and are limited to certain reference management software (e.g. EndNote) [9]. There has also been a growth in the number of fully automated tools that can deduplicate without any human involvement [10]. One limitation of these tools is that they are often tied to proprietary software and are often closed-source, meaning that the internal workings of these algorithms are largely unknown.

To address these issues around deduplication, we have designed an automation tool, the Deduplicator, available via the Systematic Review Accelerator (SRA) [11]. The Deduplicator is a free, open access, tool with a user interface that allows users to review all decisions and exports in multiple file formats allowing it to be used across different reference management software platforms. This paper has 2 objectives: (1) describe the algorithms the Deduplicator uses to detect duplicates and (2) report time and error (e.g. unique studies removed and missed duplicates) comparisons between the Deduplicator and EndNote for 10 sets of systematic review search results.

Methods

Development of the Deduplicator

Work on the Deduplicator began in June 2021, with the goal of making the deduplication of systematic review search results fast, easy and transparent. The initial design focused on replicating the semi-manual method used by the authors at the Institute for Evidence-Based Healthcare (IEBH) (i.e. using the "Find Duplicates" function in EndNote, with multiple iterations of different matches across fields). The full IEBH deduplication method is available in the supplementary materials (Supplement 1). The initial deduplication algorithm was designed on a set of five deduplicated EndNote libraries obtained from reviews published by researchers at IEBH. After internal testing on the Alpha version of the Deduplicator, the Beta version was released. In August 2021, feedback from expert information specialists was sought by emailing information and a link to the Deduplicator to the US Medical Library Association's (MLA) expertsearching email list. Feedback from multiple users was provided and incorporated into the Deduplicator. The production version of the Deduplicator was then officially released in November 2021. Since its release the Deduplicator has been accessed thousands of times.

Development of the deduplication algorithm

The initial algorithm used in the Alpha version of the Deduplicator was developed using a training dataset of five deduplicated EndNote libraries. These EndNote libraries were constructed from previous systematic reviews performed at the IEBH. These libraries were independently deduplicated manually in EndNote by two authors (JC and HG). Any differences between the two deduplicated libraries were then resolved by discussion and consensus between the authors. The development dataset is available via the IEBH/dedupe-sweep GitHub repository [12].

During development the deduplication algorithms were measured using four values:

1. True positive TP is the number of correctly identified duplicate records
2. True negative TN is the number of correctly identified unique records
3. False positive FP is the number of unique records identified as a duplicate
4. False negative FN is the number of duplicate records identified as a unique record

These values used to calculate four metrics:

1. Accuracy: provides the total number of mistakes in the deduplication process (Eq. 1)
2. Precision: provides the number of unique studies incorrectly removed in the deduplication process (Eq. 2)
3. Recall: provides the number of duplicates missed in the deduplication process (Eq. 3)
4. F1 score: combines recall and precision metrics and represents the overall performance of the model (Eq. 4)

The equations for calculating these metrics are:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1 score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{4}$$

The first algorithm ('balanced') started as a modified version of the IEBH deduplication method (Supplement 1). Following this, small modifications were iteratively made to the algorithm. These changes were then evaluated on all five libraries to evaluate if the newly modified algorithm achieves a higher accuracy/precision/recall/F1 score. Eventually, an algorithm was converged which achieved a high accuracy and precision. This algorithm was labelled the 'balanced' algorithm, and it is the algorithm that was used in the evaluation study presented in the results of this paper. After the completion of the evaluation, further improvements were made to the algorithm to optimise for either high precision or recall. This produced two improved algorithms ('relaxed' and 'focused'). The 'relaxed' algorithm is designed to minimise the number of false positives making it suitable for large libraries of records (> 2000 records) as human checking is less necessary. The 'focused' algorithm is designed to minimise the number of false negatives making it suitable for small libraries of records (< 2000 records). The results of these evaluations on the development set of libraries (without human checking) can be found in (Table 5).

Along with each algorithm, a set of mutators are specified at the top of the configuration file. These play a key role as they aim to unify differences between fields in each database. For instance, an author rewrite mutator will unify the different ways of writing author names (e.g. 'John Smith' vs 'Smith, J' vs 'J. Smith'). An alphanumeric mutator will attempt to resolve differences in Unicode characters between articles and a page number mutator will unify differences between the page numbering systems (e.g. '356-357' vs '356-7'). Unicode characters can differ across languages therefore the mutator is needed to standardise them, e.g. changing the author names Rolečková or Hammarström to Roleckova or Hammarstrom. A full table of mutators and what they do can be found in the supplementary materials (Supplement 2). These mutators are applied before deduplication and hence the process of applying all mutators will be referred to as pre-processing.

How the Deduplicator algorithm identifies duplicate records

The Deduplicator works over multiple iterations. For each iteration, multiple fields are specified, along with a primary 'sort' field which is used for the initial sort. A comparison method is also specified for each iteration (exact match or Jaro-Winkler similarity [13]). The exact match comparison method only marks a field as matching if the two strings of text match exactly. The Jaro-Winkler comparison method on the other hand returns a value between zero and one based on how closely the strings match. The algorithm works as below:

1 Apply pre-processing mutators to records to ensure they are consistently formatted (Supplement 2)
2 For each 'step' specified in the algorithm (Supplement 3):

   (a) Sort the list of records based on the specified 'sort' field (e.g. "title")
   (b) Split the records into separate sub-groups based on matching entries for the specified 'sort' field (e.g. If "title", all records with a title of "Automation of Duplicate Record Detection for Systematic Reviews" will be grouped together)
   (c) Calculate the similarity score for every combination of records inside the sub-group

3 Once all 'steps' inside the algorithm have been performed, take an average of the similarity scores calculated for each combination of records
4 If two records have an average similarity score greater than a threshold (e.g. 0.01), the two records are marked as duplicates

Using the base algorithm, deduplication algorithms can be defined in configuration files, which specify each iteration, along with what fields should be compared, what field the records should be sorted by and what comparison method to use. The full code for each deduplication method is provided in the supplementary materials (Supplement 3).

As an example, for the 'balanced' algorithm, initially the pre-processing is applied. This would include processes such as converting all title characters to lower case, removing all spaces and any non-alpha-numeric characters. Hence the title "Automation of Duplicate Record Detection for Systematic Reviews" would become "automationofduplicaterecorddetectionforsystematicreviews".

Next, the first 'step' of the algorithm specifies the 'sort' field as "title". This means that all records are sorted and then split into subgroups based on matching titles. The 'fields' for this step are specified as "title" and "volume". Because the 'comparison' is specified to be "exact", both the title and volume of the record need to exactly match to give a similarity score of 1. If any of the fields do not exactly match (including one of the fields being missing), then the similarity score will be 0.

The scores are then calculated in the same way for the four other 'steps' specified in the 'balanced' algorithm. The five scores (which were calculated at each step) are then averaged to give a final similarity score for each combination of records. If the averaged similarity score is greater than 0.01, then the two records are presumed to be duplicates.

The mean similarity score is also used to classify how likely it is that two records are duplicates. A score greater than or equal to 0.9 will put duplicate records in the "Extremely Likely Duplicates" group. A score greater than or equal to 0.7 will put duplicate records in the "Highly Likely Duplicates" group. Any score less than 0.7 but greater than 0.01 will put the duplicates in the "Likely Duplicates" group. These score thresholds are arbitrarily chosen after testing against various duplication scenarios. These scores were found to be ideal for their relative groups, such that the "Extremely Likely Duplicates" and "Highly Likely Duplicates" groups are very unlikely to contain any unique records (false positives).

Further information and the code for the algorithm is available via the IEBH/dedupe-sweep GitHub repository [12].

Evaluation of the Deduplicator

The Deduplicator was evaluated by two screeners (HG and JC) using search results from a set of 10 randomly selected Cochrane reviews. To avoid any confounding from a learning effect, we used a cross-over, paired design where person one would deduplicate the search results using EndNote, while person two would deduplicate using the Deduplicator. They would then switch methods, so person one would deduplicate the next set of search results using the Deduplicator and person two would deduplicate using EndNote. The time taken to deduplicate the search results and the numbers of removed unique studies and missed duplicates were compared.

Definition of a duplicate record

There is currently a lack of an agreed upon definition of what is a duplicate record. For our study we have defined a duplicate as the same article published in the same place, while the same article published in a different place is not a duplicate. An example of this is the PRISMA statement which was published in multiple journals.

These are duplicates:

- Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. J Clin Epidemiol. 2009 Oct;62(10):1006-12. doi: 10.1016/j.jclinepi.2009 .06.005
- Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G. (2009). Journal of Clinical Epidemiology, 62(10), 1006-1012. https://doi.org/10.1016/j.jclinepi.2009.06.005
- Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. J Clin Epidemiol. 2009;62(10):1006-1012. doi:10.1016/j.jclinepi.2009.06.005

These are not duplicates:

- Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group.Int J Surg. 2010;8(5):336-41. doi: 10.1016/j.ijsu.2010.02.007
- Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group.J Clin Epidemiol. 2009 Oct;62(10):1006-12. doi: 10.1016/j.jclinepi.2009 .06.005
- Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. BMJ. 2009 Jul 21;339:b2535. doi: 10.1136/bmj.b2535

Selection of systematic reviews to be deduplicated

To ensure an unbiased sample of search results to be used, we randomly selected 10 Cochrane reviews published in the last 5 years (January 2017–September 2021). To randomly select the systematic reviews, the following search string was run in PubMed; "Cochrane Database Syst Rev[Journal] AND 2017:2021[pdat]". Then, a random number was generated using the Google random number generator. This number was between one and the total number of search results found (e.g. if 5000 results were found, the random number was set to be between one and 5000). The search result that then corresponded to the random number generated was checked to ensure it meets the inclusion criteria. This continued until 10 Cochrane reviews were identified.

Inclusion criteria of the systematic reviews

To be selected and used in the study, the search strategy in the Cochrane review had to meet the following criteria:

- All search strings for all databases needed to be reported in the review
- The number of databases searched in the review had to be two or more
- The total number of search results found by the combination of all search strings had to be between 500 and 10,000 records

The decision to limit search results to 500 to 10,000 was to reduce variability between samples to be deduplicated and to ensure they were representative of a typical systematic review which have a median size of 1781 records [1].

Obtaining the sample to be deduplicated

After 10 eligible systematic reviews were selected, their searches for all bibliographic databases were run and the results exported and collated in EndNote. No date or language limits were applied, and searches of specialised registers, trial registries and grey literature were excluded.

Deduplication of search results

Two screeners (HG and JC) independently deduplicated 10 sets of search results. HG is a research assistant (now PhD candidate) with 2 years' experience with systematic reviews but with no experience deduplicating search results. JC is an information specialist with over 15 years' experience with systematic reviews and deduplicating. HG screened the odd numbered sets of search results using EndNote (1, 3, 5, 7 and 9) then screened the even numbered sets with the Deduplicator (2, 4, 6, 8 and 10). JC screened the even numbered sets with EndNote (2, 4, 6, 8 and 10) and the odd numbered sets with the Deduplicator (1, 3, 5, 7 and 9) (Table 1). EndNote deduplication is defined as using the IEBH EndNote deduplication method (Supplement 1), while Deduplicator is the solution discussed in this paper. In the Deduplicator, the Beta algorithm (referred to as the 'balanced' algorithm) was used.

Validation of deduplication

To identify errors (i.e. a duplicate mistakenly marked as non-duplicate, and vice versa), the screener's libraries were compared. This was done once all 10 sample sets had been deduplicated. Any discrepancies were manually checked and verified by consensus between two authors (HG and CF). This produced a final "correctly deduplicated" EndNote library for each sample set. This enabled the identification of errors from each screeners' library, with an incorrectly removed unique article labelled a "false positive", while a duplicate which was incorrectly missed was labelled as a "false negative".

Outcomes

We evaluated the Deduplicator by four outcomes:

1 *Time required to deduplicate*: each screener recorded how long it took to perform deduplication on each library in minutes using a phone timer. The screener started the timer from when the file was first open and stopped the timer when they were satisfied that all duplicates were identified
2 *Unique studies removed/False positives*: the number of records in the library the screener classified as a duplicate when they were a unique record
3 *Duplicates missed/False negatives*: the number of records in the library the screener classified as a unique record when it was a duplicate record
4 *Total errors*: (false positives + false negatives)

Comparison between Deduplicator algorithms

In addition to testing the five development libraries against each Deduplicator algorithm ('balanced', 'focused' and 'relaxed'), we also performed an additional head-to-head evaluation between the three Deduplicator algorithms taking a dataset from a previous deduplication study by Rathbone et al. [24]. This dataset contains four sets of search results from studies related to: cytology-screening, haematology, respiratory and stroke. The full breakdown of the dataset is provided in Table 2. All

Table 1 Assignment of EndNote vs Deduplicator methods between researchers

Set no.	Systematic review (author year)	Number of records	Hannah Greenwood	Justin Clark
1	Lorentzen 2020 [14]	813	EndNote	Deduplicator
2	Alebed 2020 [15]	1479	Deduplicator	EndNote
3	Dawson 2021 [16]	3912	EndNote	Deduplicator
4	Wiffen 2017 [17]	1028	Deduplicator	EndNote
5	Kamath 2020 [18]	1785	EndNote	Deduplicator
6	Ghobara 2017 [19]	1807	Deduplicator	EndNote
7	Bennett 2018 [20]	2111	EndNote	Deduplicator
8	Hannon 2021 [21]	1061	Deduplicator	EndNote
9	Roberts 2020 [22]	3181	EndNote	Deduplicator
10	Jaschinski 2018 [23]	2447	Deduplicator	EndNote

three algorithms were run as is, meaning that there was no manual checking by a human as there was in the EndNote comparison.

Like with the development libraries, accuracy, precision, recall and F1 score were the four measures used for comparison between the Deduplicator algorithms. A high precision score indicates that few unique studies were identified as duplicates. A high recall score indicates that very few duplicate studies were incorrectly classified as unique studies. F1 score is a combination score of both precision and recall. The formula for these measures are presented in Eqs. 2, 3 and 4.

Results

Time taken to deduplicate

The mean size of the sample sets was 1962 records (range: 813 to 3912). The mean time required to deduplicate the sample sets with the Deduplicator was 8 min (range: 4 to 20 min) compared to a mean time of 27 min (range 6 to 76 min) using the semi-manual EndNote method. This equates to a mean time reduction of 67% (19 min) when deduplicating search results (Fig. 1).

Table 2 Breakdown of dataset used for comparison of Deduplicator algorithms [24]

Study	Number of records	Number of duplicates	Number of unique studies
Cytology screening	1856	1404	452
Haematology	1415	246	1169
Respiratory	1988	799	1189
Stroke	1292	507	785

Number of errors

The mean number of errors when using the Deduplicator was 3.3 (range: 0 to 7), while the mean number of errors when using EndNote was 6.2 (range: 0 to 16). The mean error rate for screeners using Deduplicator was 47% less compared to EndNote (Table 3).

The mean number of unique studies removed was 1.5 (range: 0 to 3) with the Deduplicator and 3.3 (range: 0 to 12) with EndNote. The mean number of duplicates missed was 1.8 (range: 0 to 5) with the Deduplicator and 2.9 (range: 0 to 8) with EndNote (Table 3)

Normalised time and error rates

In order to reduce the bias of large libraries on the mean measurements for time and error rate, here we normalise each of the systematic reviews to be measured per 1000 records deduplicated. The mean time to deduplicate 1000 records was 5 min with Deduplicator compared to 15 min with EndNote (Table 4). The mean time to deduplicate 1000 records using Deduplicator is 67% less than EndNote. The mean number of errors per 1000 records was 1.8 with Deduplcicator compared to 3.1 with EndNote (Table 4). The mean number of errors per 1000 records is 42% less with Deduplicator compared to EndNote.

Analysis between screeners

All measurements in this section are normalised to be measured per 1000 records deduplicated, to negate the difference in mean library size between screeners. The mean time for the experienced screener (JC) was 3 min/1000 records (range: 2 to 5 min) using the Deduplicator and 9 min/1000 records (range: 6 to 13 min) using EndNote. The mean time for the inexperienced screener

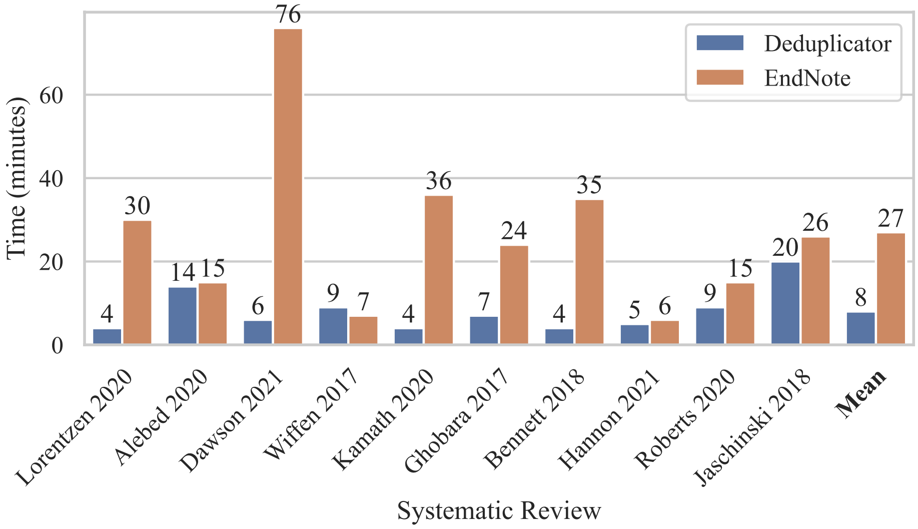

Fig. 1 Time taken to deduplicate every systematic review with each method

Table 3 Comparison of number of errors for each library

Systematic review	Total records	Deduplicator			EndNote		
		Unique studies removed	**Duplicates missed**	**Total errors**	**Unique studies removed**	**Duplicates missed**	**Total errors**
Lorentzen 2020	813	0	0	0	1	1	2
Alebed 2020	1479	1	5	6	5	3	8
Dawson 2021	3912	2	0	2	2	5	7
Wiffen 2017	1028	1	0	1	0	0	0
Kamath 2020	1785	0	2	2	1	1	2
Ghobara 2017	1807	2	4	6	3	2	5
Bennett 2018	2111	1	2	3	2	2	4
Hannon 2021	1061	3	0	3	2	3	5
Roberts 2020	3181	3	0	3	12	4	16
Jaschinski 2018	2447	2	5	7	5	8	13
Mean	**1962.4**	**1.5**	**1.8**	**3.3**	**3.3**	**2.9**	**6.2**

Table 4 Time to deduplicate and error rate per 1000 records

	Time per 1000 records (minutes)		Total errors per 1000 records	
Systematic review	**Deduplicator**	**EndNote**	**Deduplicator**	**EndNote**
Lorentzen 2020	5	37	0.0	2.5
Alebed 2020	10	10	4.1	5.4
Dawson 2021	2	19	0.5	1.8
Wiffen 2017	9	7	1.0	0.0
Kamath 2020	2	20	1.1	1.1
Ghobara 2017	4	13	3.3	2.8
Bennett 2018	2	17	1.4	1.9
Hannon 2021	5	6	2.8	4.7
Roberts 2020	3	5	0.9	5.0
Jaschinski 2018	8	11	2.9	5.3
Mean	**5**	**15**	**1.8**	**3.1**

(HG) was 7 min/1000 records (range: 4 to 9 min) using the Deduplicator and 20 min/1000 records (range: 5 to 37 min) using EndNote (Fig. 2).

The experienced systematic reviewer (JC) when using the Deduplicator had a mean error rate of 0.8 per 1000 records. Using EndNote, JC had a mean error rate of 3.6 per 1000 records (Fig. 3). The inexperienced systematic reviewer (HG) when using the Deduplicator had a mean error rate of 2.8 per 1000 records. When using EndNote, HG had a mean error rate of 2.5 per 1000 records (Fig. 3).

Comparison between Deduplicator algorithms

Testing against the 5 development libraries of records showed the focused algorithm achieved the highest mean recall of 0.9999 and the highest overall F1 score of 0.9966. The 'relaxed' algorithm achieved the highest mean precision of 0.9996 (Table 5).

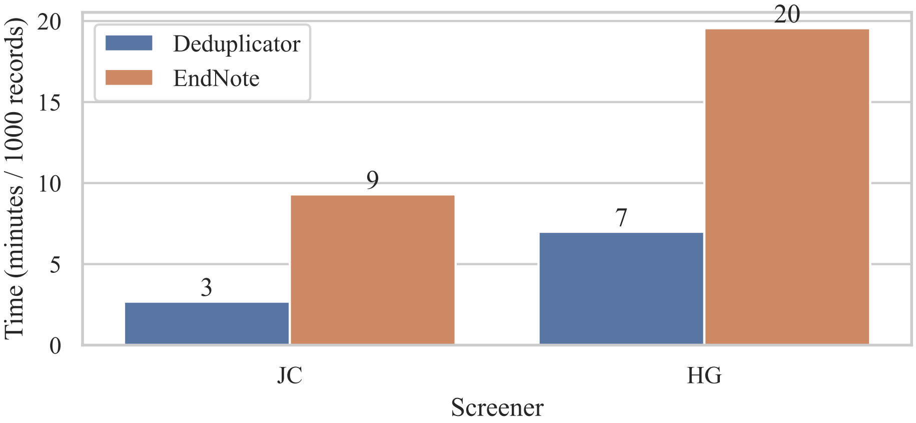

Fig. 2 Mean time taken for each screener to deduplicate 1000 records

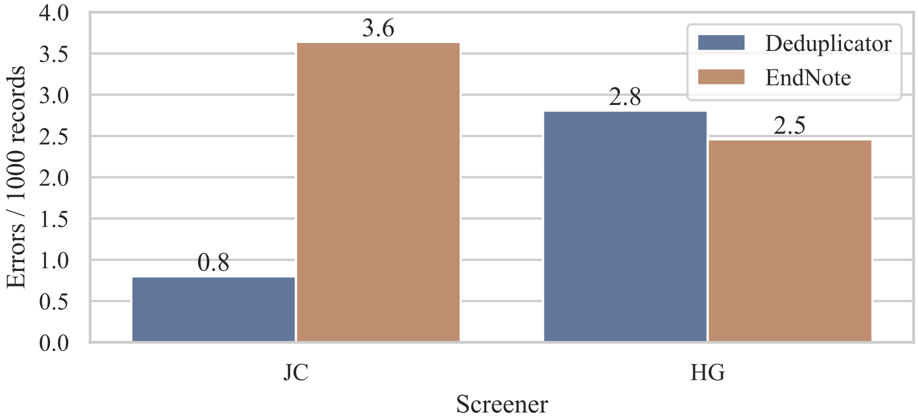

Fig. 3 Mean number of total errors per 1000 records for each screener

Expanding this analysis to the unseen dataset taken from the study performed by Rathbone et al. [24] shows similar results. The 'balanced' algorithm had the highest mean F1 score of 0.9647, although 'focused' is not far behind at 0.9645. 'Focused' has the highest mean recall of 0.9757 while the 'relaxed' algorithm has the highest mean precision of 0.9896 (Table 6).

Discussion

After the development and validation of the Deduplicator, we conducted a study to compare Deduplicator to a manual EndNote method on outcomes of time taken to deduplicate and number of errors made. We found the Deduplicator reduced the mean time needed to deduplicate by approximately 67%, from 15 min per 1000 records with EndNote to 5 min with Deduplicator (Table 4). We also found that fewer mistakes were made, with a mean error reduction of approximately 42%, from 3.1 errors per 1000 records with EndNote to 1.8 with Deduplicator (Table 4). Although this was only a small study (with two participants and 10 sets of search results deduplicated), it provides preliminary evidence that using the Deduplicator is superior to the widely-used method of deduplicating using EndNote, on outcomes of time and error rate.

Table 5 Accuracy, precision, recall and F1 score for each of the Deduplicator algorithms on the development libraries

Algorithm	Study	Accuracy	Precision	Recall	F1 score
Balanced	Blue light	0.9989	1.0000	0.9979	0.9990
Balanced	Copper	0.9822	0.9892	0.9786	0.9839
Balanced	Diabetes	0.9909	0.9890	0.9919	0.9904
Balanced	Tafenoquine	0.9888	1.0000	0.9825	0.9912
Balanced	UTI	1.0000	1.0000	1.0000	1.0000
Focused	Blue light	1.0000	1.0000	1.0000	1.0000
Focused	Copper	0.9941	0.9894	1.0000	0.9947
Focused	Diabetes	0.9913	0.9823	0.9997	0.9909
Focused	Tafenoquine	1.0000	1.0000	1.0000	1.0000
Focused	UTI	0.9981	0.9950	1.0000	0.9975
Relaxed	Blue light	0.9977	1.0000	0.9958	0.9979
Relaxed	Copper	0.9921	1.0000	0.9858	0.9928
Relaxed	Diabetes	0.9934	0.9982	0.9878	0.9930
Relaxed	Tafenoquine	0.9944	1.0000	0.9912	0.9956
Relaxed	UTI	0.9799	1.0000	0.9475	0.9730
Balanced	**Mean**	0.9921	0.9956	0.9902	0.9929
Focused	**Mean**	**0.9967**	0.9934	**0.9999**	**0.9966**
Relaxed	**Mean**	0.9915	**0.9996**	0.9816	0.9905

When using the Deduplicator, the error rates for JC were substantially lower compared to HG with 0.8 errors vs 2.8 errors per 1000 records respectively (Fig. 3). One explanation for this is the difference in experience levels between the screeners. One of the screeners (HG) is new to systematic reviews and had minimal experience deduplicating search results, while the other (JC) has years' of experience and has deduplicated many sets of search results. This may facilitate JC to be better at accurately spotting duplicates compared to HG. However, for the EndNote deduplication method, HG had a lower error rate compared to JC with 2.5 vs 3.6 errors per 1000 records respectively (Fig. 3). This may be explained by the extra time that HG took when deduplicating using EndNote compared to JC, where HG took 20 min per 1000 records compared to 9 for JC (Fig. 1). The error rate for Deduplicator and EndNote were similar for HG, however Deduplicator facilitated much faster screening for HG, reducing the time to screen from 20 min per 1000 records to 7 min per 1000 records (Fig. 1).

After the evaluation it became clear that the 'balanced' algorithm could be improved upon. Also, as usage of the Deduplicator increased, two different use cases emerged. There were users who wanted to duplicate libraries of records without any manual check and those who wanted to be able to check each decision made by the Deduplicator. This led to the development of two algorithms, 'relaxed' and 'focused' which replaced the 'balanced' algorithm. When comparing algorithms, the 'focused' algorithm had the highest recall score, indicating it was the best at finding all duplicates; however, it has the lowest precision score which means that the results need to be checked. The 'relaxed' algorithm had the highest precision, meaning it is unlikely to remove any unique studies; however, it has the lowest recall meaning that some duplicate studies will remain after deduplication (Tables 5 and 6). Therefore, we recommend the 'relaxed' algorithm for large libraries of records (> 2000 records), where people do not wish to check the results and the 'focused' algorithm for small libraries of records (< 2000 records) as this is a feasible number to check manually. These numbers may change depending on the time constraints of the individual study.

In addition to the tools investigated here (EndNote and Deduplicator), there are multiple other tools to help with deduplication. Generally, they are built into database platforms (e.g. Ovid or EBSCO), reference management software (e.g. EndNote, Mendeley or Zotero) or systematic review software (e.g. Rayyan or Covidence). The primary advantage of Deduplicator compared to other tools listed here is that it is fully open-source, free to use and not connected to any existing reference management software or database platforms. However, unlike some tools such as Covidence, Deduplicator requires exporting the library from a reference manager and then importing the result back into the reference manager or screening tool to continue with screening. While this is something that is being worked on, some users may find it undesirable to move their records between different tools.

Table 6 Accuracy, precision, recall and F1 score for each of the Deduplicator algorithms on the Rathbone et al. dataset [24]

Algorithm	Study	Accuracy	Precision	Recall	F1 score
Balanced	Cytology-screening	0.9758	0.9836	0.9843	0.9840
Balanced	Haematology	0.9696	0.9177	0.9065	0.9121
Balanced	Respiratory	0.9819	0.9823	0.9725	0.9774
Balanced	Stroke	0.9884	0.9824	0.9882	0.9853
Focused	Cytology-screening	0.9790	0.9789	0.9936	0.9862
Focused	Haematology	0.9654	0.8774	0.9309	0.9034
Focused	Respiratory	0.9864	0.9801	0.9862	0.9832
Focused	Stroke	0.9884	0.9786	0.9921	0.9853
Relaxed	Cytology-screening	0.9763	0.9885	0.9801	0.9843
Relaxed	Haematology	0.9710	0.9812	0.8496	0.9107
Relaxed	Respiratory	0.9779	0.9948	0.9499	0.9718
Relaxed	Stroke	0.9853	0.9939	0.9684	0.9810
Balanced	**Mean**	0.9789	0.9665	0.9629	**0.9647**
Focused	**Mean**	**0.9798**	0.9538	**0.9757**	0.9645
Relaxed	**Mean**	0.9776	**0.9896**	0.9370	0.9619

A study conducted by Guimarães et al. [25] evaluated five different tools for deduplication: EndNote X9, Mendeley, Zotero, Rayyan and the Deduplicator (listed in the study as SRA). The results of this study found that specificity, or the proportion of non-duplicates correctly identified as such, was best in Mendeley and the Deduplicator, with both achieving a 1.00 score. It also found that sensitivity, or the ability to correctly identify duplicates, was highest for Rayyan, Mendeley and the Deduplicator. The study found that Rayyan had 35.1 errors per 1000 records, Zotero had 23.8, EndNote had 17.7, Mendeley had 3.3 and the Deduplicator had 2.5 errors per 1000 records. This study suggests that Deduplicator has the lowest error rate and is consistent with the results obtained from our study of 1.8 errors per 1000 records (Table 4).

Another study published by McKeown et al. [10] evaluated some other commonly used deduplication tools. The study found that the number of errors was lowest when using the Ovid database platform, with 90 errors (28.8 per 1000 records). This is not suitable for most reviewers as it requires all searches to be run in Ovid databases (e.g. if you use PubMed or CINAHL this method is not usable). The systematic review software performed next best with Rayyan having 101 errors (32.3 per 1000 records) and Covidence with 122 errors (39.0 per 1000 records). Finally, the reference management software performed worst with Mendeley having 212 errors (67.7 per 1000 records), Zotero having 619 (197.8 per 1000 records) and EndNote having 739 (236.1 per 1000 records). However, the results for the EndNote method from this study can't be directly compared to our results as their study used the default EndNote algorithm with no manual human check.

It is also worth mentioning another recent deduplication tool, "Deduklick". In research conducted by Borissov et al. [26] Deduklick achieved an impressive mean recall of 99.51% with 100% precision. While our study design does not allow for direct precision or recall calculations, it would be worthwhile for future comparative research to investigate performance of Deduklick vs other deduplication methods.

Limitations

One of the limitations of the study is the discrepancy in experience between the two authors. For example, in the “Wiffen, 2017” systematic review, the Deduplicator was slightly slower to deduplicate [HG] compared to the semi-manual EndNote method [JC]. JC’s extra experience probably facilitated quick, accurate semi-manual deduplication of the small Wiffen library faster than HG could achieve using the Deduplicator. This difference in deduplication speed/accuracy between authors is partially mitigated by the equal split of methods used by each author, but this does not eliminate this bias entirely. Despite this disparity, using the Deduplicator increased the speed with which both screeners could deduplicate sets of search results (Fig. 2). It could also be argued that Deduplicator will likely be used by researchers with a broad range of experience, and therefore having two types of screener experience level in this study makes it more representative of real world conditions.

A second limitation is the possibility that both authors made the same mistake, e.g. both missed the same duplicate record. This error would not show up in the results, as the errors were determined by comparing both screeners’ results. But, since deduplication was done separately by two people with the aid of a computer algorithm, we can be fairly confident this number is low. Also, as this is a comparison to determine which deduplication method was better, if neither had the error marked against them, this would not affect the comparison in errors made between the two methods.

Third, only the Beta, or ‘balanced’ algorithm was assessed in the direct comparison to EndNote. Since the completion of the study, the ‘balanced’ algorithm has been replaced by two new algorithms: the ‘relaxed’ and ‘focused’ algorithms. While these were not compared directly against EndNote, they were compared against the ‘balanced’ algorithm. The results for this analysis is presented in Table 6.

Fourth, as this is an efficacy trial using selected datasets, the real-world time-saving and error rate of the Deduplicator still needs to be evaluated.

Future research

Future work in this area will need to focus on two main areas, comparing the newest version of the Deduplicator to other deduplication tools on common datasets and performance in real world settings. Due to the difference in data, we could not directly compare our results to those reported in other studies, such as the study by Mckeown [10]. Therefore, plans are currently underway to collate a new set of search results, with all duplicates detected, to be used in a comparative study of all known and available deduplication tools. Once this second, experimental, study is complete, planning will begin to determine the effectiveness of the Deduplicator in a real-world setting.

Conclusion

This study demonstrates that using the Deduplicator for duplicate record detection reduces the time taken and errors made when compared to using a semi-manual EndNote method. The Deduplicator also allows an easier point of entry for new researchers to begin deduplicating, and it compares favourably with the error rates of other tools and methods.

Abbreviations

IEBH	Institute for Evidence-Based Healthcare
SRA	Systematic Review Accelerator

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13643-024-02619-9.

Additional file 1. A PDF with a complete guide to the IEBH deduplication process using EndNote.

Additional file 2. A PDF file with a table for the various Deduplicator mutators and a brief description of each.

Additional file 3. A PDF file representing the JSON code that is used for the deduplication comparison algorithm.

Acknowledgements

Dr. Henry Nguyen (Griffith University) provided general support and advice during writing of the manuscript.

Authors’ contributions

CF contributed code for the Deduplicator algorithm and was the primary contributor in writing the manuscript and interpreting the data. JC is responsible for the conception of the study design as well as the original draft of the manuscript abstract. JC and HG performed deduplication of the experimental libraries as well as providing substantial revisions to the manuscript. MC is responsible for the conception of the original Deduplicator algorithm, contributed to the code for Deduplicator and provided substantial revisions to the manuscript. All authors read and approved the final manuscript.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

Availability of data and materials

The full code for Deduplicator including the development library datasets are available via the IEBH/dedupe-sweep GitHub repository [12]. The data that support the findings of this study are available from the corresponding author, CF, upon reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors of this study are responsible for the development of the Deduplicator and hence may present bias towards favourable findings. However, we encourage independent testing of the method and have made the code and testing datasets open-source and publicly available to be as transparent as possible and improve replicability.

Received: 22 June 2023 Accepted: 18 July 2024
Published online: 02 August 2024

References

1. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open. 2017;7(2):e012545. https://doi.org/10.1136/bmjopen-2016-012545.
2. Murad MH, Asi N, Alsawas M, Alahdab F. New evidence pyramid. BMJ Evid-Based Med. 2016;21(4):125–7. https://doi.org/10.1136/ebmed-2016-110401.
3. Michelson M, Reuter K. The significant cost of systematic reviews and meta-analyses: a call for greater involvement of machine learning to assess the promise of clinical trials. Contemp Clin Trials Commun. 2019;16:100443. https://doi.org/10.1016/j.conctc.2019.100443.
4. Scott AM, Glasziou P, Clark J. We extended the 2-week systematic review (2weekSR) methodology to larger, more complex systematic reviews: a case series. J Clin Epidemiol. 2023;157:112–9. https://doi.org/10.1016/j.jclinepi.2023.03.007.
5. Tufanaru C, Surian D, Scott AM, Glasziou P, Coiera E. The 2-week systematic review (2weekSR) method was successfully blind-replicated by another team: a case study. J Clin Epidemiol. 2024;165. https://doi.org/10.1016/j.jclinepi.2023.10.013.
6. Beller E, Clark J, Tsafnat G, Adams C, Diehl H, Lund H, et al. Making progress with the automation of systematic reviews: principles of the International Collaboration for the Automation of Systematic Reviews (ICASR). Syst Rev. 2018;7(1):77. https://doi.org/10.1186/s13643-018-0740-7.
7. Tsafnat G, Glasziou P, Choong MK, Dunn A, Galgani F, Coiera E. Systematic review automation technologies Syst Rev. 2014;3:74. https://doi.org/10.1186/2046-4053-3-74.
8. Qi X, Yang M, Ren W, Jia J, Wang J, Han G, et al. Find duplicates among the PubMed, EMBASE, and Cochrane Library Databases in systematic review. PLoS ONE. 2013;8(8):e71838. https://doi.org/10.1371/journal.pone.0071838.
9. Bramer WM, Giustini D, de Jonge GB, Holland L, Bekhuis T. De-duplication of database search results for systematic reviews in EndNote. J Med Libr Assoc. 2016;104(3):240–3. https://doi.org/10.3163/1536-5050.104.3.014.
10. 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. https://doi.org/10.1186/s13643-021-01583-y.
11. IEBH. The Systematic Review Accelerator. 2018. https://sr-accelerator.com. Accessed 11 Nov 2022.
12. IEBH. Deduplicator GitHub Repository. 2020. https://github.com/IEBH/dedupe-sweep. Accessed 11 Nov 2022.
13. Winkler W. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. Proceedings of the Section on Survey Research Methods. American Statistical Association. Alexandri: American Statistical Association; 1990. Avaliable at: https://eric.ed.gov/?id=ED325505.
14. Lorentzen AK, Davis C, Penninga L. Interventions for frostbite injuries. Cochrane Database Syst Rev. 2020;12(12):CD012980. https://doi.org/10.1002/14651858.CD012980.pub2.
15. Alabed S, Sabouni A, Al Dakhoul S, Bdaiwi Y. Beta-blockers for congestive heart failure in children. Cochrane Database Syst Rev. 2020;7(7):CD007037. https://doi.org/10.1002/14651858.CD007037.pub4.
16. Dawson JA, Summan R, Badawi N, Foster JP. Push versus gravity for intermittent bolus gavage tube feeding of preterm and low birth weight infants. Cochrane Database Syst Rev. 2021;8(8):CD005249. https://doi.org/10.1002/14651858.CD005249.pub3.
17. Wiffen PJ, Cooper TE, Anderson AK, Gray AL, Grégoire MC, Ljungman G, et al. Opioids for cancer-related pain in children and adolescents. Cochrane Database Syst Rev. 2017;7(7):CD012564. https://doi.org/10.1002/14651858.CD012564.pub2.
18. Kamath MS, Mascarenhas M, Kirubakaran R, Bhattacharya S. Number of embryos for transfer following in vitro fertilisation or intra-cytoplasmic sperm injection. Cochrane Database Syst Rev. 2020;8(8):CD003416. https://doi.org/10.1002/14651858.CD003416.pub5.
19. Ghobara T, Gelbaya TA, Ayeleke RO. Cycle regimens for frozen-thawed embryo transfer. Cochrane Database Syst Rev. 2017;7(7):CD003414. https://doi.org/10.1002/14651858.CD003414.pub3.
20. Bennett MH, Feldmeier J, Smee R, Milross C. Hyperbaric oxygenation for tumour sensitisation to radiotherapy. Cochrane Database Syst Rev. 2018;4(4):CD005007. https://doi.org/10.1002/14651858.CD005007.pub4.
21. Hannon CW, McCourt C, Lima HC, Chen S, Bennett C. Interventions for cutaneous disease in systemic lupus erythematosus. Cochrane Database Syst Rev. 2021;3(3):CD007478. https://doi.org/10.1002/14651858.CD007478.pub2.
22. Roberts KE, Rickett K, Feng S, Vagenas D, Woodward NE. Exercise therapies for preventing or treating aromatase inhibitor-induced musculoskeletal symptoms in early breast cancer. Cochrane Database Syst Rev. 2020;1(1):CD012988. https://doi.org/10.1002/14651858.CD012988.pub2.
23. Jaschinski T, Mosch CG, Eikermann M, Neugebauer EA, Sauerland S. Laparoscopic versus open surgery for suspected appendicitis. Cochrane Database Syst Rev. 2018;11(11):CD001546. https://doi.org/10.1002/14651858.CD001546.pub4.
24. Rathbone J, Carter M, Hoffmann T, Glasziou P. Better duplicate detection for systematic reviewers: evaluation of Systematic Review Assistant-Deduplication Module. Syst Rev. 2015;4(1). https://doi.org/10.1186/2046-4053-4-6.
25. Guimarães NS, Ferreira AJF, Ribeiro Silva RdC, de Paula AA, Lisboa CS, Magno L, et al. Deduplicating records in systematic reviews: there are free, accurate automated ways to do so. J Clin Epidemiol. 2022;152:110–115. https://doi.org/10.1016/j.jclinepi.2022.10.009.
26. 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. https://doi.org/10.1186/s13643-022-02045-9.

Publisher's Note