RESEARCH REPORT

Identifying and removing duplicate records from systematic review searches*

Yoojin Kwon, MLIS; Michelle Lemieux, MLIS; Jill McTavish, PhD, MLIS; Nadine Wathen, PhD

See end of article for authors' affiliations

DOI: http://dx.doi.org/10.3163/1536-5050.103.4.004

Objective: The purpose of this study was to compare effectiveness of different options for de-duplicating records retrieved from systematic review searches.

Methods: Using the records from a published systematic review, five de-duplication options were compared. The time taken to de-duplicate in each option and the number of false positives (were deleted but should not have been) and false negatives (should have been deleted but were not) were recorded.

Results: The time for each option varied. The number of positive and false duplicates returned from each option also varied greatly.

Conclusion: The authors recommend different de-duplication options based on the skill level of the searcher and the purpose of de-duplication efforts.

Keywords: Biomedical Research/Standards, Duplicate Publication as Topic, Publications/Standards, Review Literature as Topic

INTRODUCTION

Systematic reviews continue to gain prevalence in health care primarily because they summarize and appraise vast amounts of evidence for busy health care providers [1, 2]. Because they are used as the foundation for clinical and policy-related decision-making processes, it is critical to ensure that the methods used in systematic reviews are explicit and valid. The Cochrane Collaboration, for example, places a heavy emphasis on minimizing bias with a thorough, objective, and reproducible multi-database search [2], which has become the standard in systematic review processes [3]. Searching multiple databases, however, results in the retrieval of numerous duplicate citations. Also, due to the nature of the publishing cycle in the field of medicine, conference abstracts and full-text articles reporting the same information are often retrieved concurrently. In addition, although many have called out against such practice, some authors "slice, reformat, or reproduce material from a study" [4], which creates repetitive, duplicate, and redundant publications. As Kassirer and Angell argued, "multiple reports of the same observations can over emphasize the importance of the findings, overburden busy reviewers, fill the medical literature with inconsequential material, and distort the academic reward system" [5]. Removing these duplicate citations, also known as de-duplication, can be a time-consuming process but is necessary to ensure a valid and reliable pool of studies for inclusion in a systematic review.

The aim of this study was to explore and compare the effectiveness of various de-duplication features. Specifically, the authors examined and compared two categories of de-duplication strategies: de-duplicating in the Ovid and EBSCO database

* Based on a poster session at Canadian Health Libraries Association/Association des bibliothéques de la santé du Canada (CHLA/ABSC) '15, Riding the Wave of Change; Vancouver, BC; June 20, 2015.

This article has been approved for the Medical Library Association's Independent Reading Program <http://www.mlanet.org/education/irp/>.

A supplemental appendix and supplemental Table 1 and Table 3 are available with the online version of this journal.

platforms and de-duplicating in three selected reference management software packages: RefWorks, EndNote, and Mendeley.

METHODS

Five de-duplication options were examined in this study:

1. Ovid multifile search: Searchers are able to de-duplicate in the Ovid platform across various Ovid products, such as Ovid MEDLINE and Ovid Embase.
2. CINAHL (EBSCO) and Ovid multifile search: Searchers are able to exclude MEDLINE records in the CINAHL database.
3. Refworks: Searchers are able to de-duplicate all records from various sources in this citation manager.
4. Mendeley: This citation manager automatically identifies duplicates among imported references, which can be deleted.
5. Endnote: When de-duplicating, this citation manager creates a separate group for duplicate references only. It is possible for searchers to view this group and delete the duplicates.

To create the citation samples used for this study, we reran the search strategies that were developed for a systematic review on ward closure as an infection control practice in Ovid MEDLINE, Ovid Embase, and CINAHL from the database inception to September 11, 2014 (Appendix, online only) [6].

For the Ovid multifile option (option 1), which allows de-duplication across various Ovid products, we opened up MEDLINE and Embase in the Ovid platform and ran a search using the strategies that were designed for the aforementioned systematic review. We ran the "use" command and database codes for MEDLINE and Embase, which are "pmoz" and "oemezd," respectively, to ensure that the retrieved results were filtered appropriately (Appendix, online only). Then, we used the "remove duplicates" command for de-duplication.

For the EBSCO CINAHL option (option 2), we ran a search in CINAHL and limited the search results to non-MEDLINE citations. The results from the searches in Ovid and EBSCO were collated and recorded in two spreadsheets: the first one contained Ovid results only, and the second one contained both Ovid and EBSCO results.

For the other three options (RefWorks, Endnote, and Mendeley), we retrieved all citations from the systematic review and exported them to each de-duplication option. In RefWorks, we clicked on the "Exact Duplicates" and "Close Duplicates" buttons in the "View" tab and deleted all identified citations. In EndNote, we clicked on the "Find Duplicates" button under the "References" menu. We deleted everything in the EndNote library duplicate references group. We loaded references as a Research Information Systems (RIS) file into Mendeley, where they were automatically de-duplicated. "Check duplicates" from the tools menu was then run to check for close duplicates, all of which were merged. All sets of citations were downloaded and recorded on separate spreadsheets.

To investigate these five de-duplication options, we needed a sample set of citations and a "gold standard" file of de-duplicated references to compare against each option. To create the sample set of citations for this study, we reran search strategies that were developed for a systematic review on ward closure as an infection control practice in Ovid MEDLINE, Ovid Embase, and CINAHL from the database inception to September 11, 2014 [6]. All of these search strategies are provided in the online appendix.

To develop the gold standard sets, we screened and de-duplicated the citations by hand, which were recorded on a Microsoft Excel spreadsheet. The detailed steps that we took to identify the duplicates in Excel are listed in the online appendix. To be considered duplicates, two or more citations had to share the same author, title, publication date, volume, issue, and start page information. The full-text versions of the citations were consulted when we were in doubt. In such cases, we also checked the population sizes, methodology, and outcomes to determine whether the citations were duplicates. Conference abstracts were deemed to be duplicates if full-text articles that shared the same study design, sample size, and conclusion were retrieved, even if their publication dates varied. Older versions of systematic reviews were deleted when there was a link between them and newer versions. All citations that were classified as duplicates were deleted from the spreadsheet. Ultimately, 2 gold standard sets were developed: one for just Ovid MEDLINE and Ovid Embase (1,087 citations) and the other for Ovid MEDLINE, Ovid Embase, and CINAHL (1,262 citations). The first gold standard set was developed for comparison against the results from the Ovid multifile search alone (option 1). The second gold standard set was developed for comparison against the other 4 options (options 2–5).

	No. before de-duplication	No. after de-duplication	Gold standard citations	False negatives†	False positives‡
Ovid*	1,253	1,178	1,087	91	0
Ovid and EBSCO	2,181	1,315	1,262	96	43
RefWorks	2,181	1,353	1,262	94	3
EndNote	2,181	1,514	1,262	258	6
Mendeley	2,181	1,294	1,262	36	4

* Compared against the gold standard set for Ovid MEDLINE and Ovid Embase only.
† Duplicate citations that should have been deleted but were not.
‡ Duplicate citations that were deleted but should not have been.

Table 2
Number of de-duplicated citations and breakdown

All sets of results from the de-duplication strategies outlined above were compared against the gold standard sets to identify false negatives (duplicate citations that should have been deleted but were not) and false positives (duplicate citations that were deleted but should not have been). We also recorded the time it took to de-duplicate results in each option (Table 1, online only). We took into consideration the results of this comparison and the time it took to de-duplicate with each option when determining the most effective strategy for de-duplication when searching the selected databases and using the selected reference management software.

RESULTS

The time spent on each de-duplication option varied (Table 1, online only). Including the time spent on reaching consensus, developing the gold standard samples of non-duplicate results took four hours and forty-five minutes. Carrying out Ovid multifile and CINAHL searches took less than three minutes to retrieve the results. Likewise, the Ovid multifile and CINAHL non-MEDLINE searches each took under three minutes. RefWorks took approximately ten minutes to delete exact and close duplicates. EndNote took three minutes to load and delete duplicates. Mendeley took five minutes. The majority of this time was spent merging the close duplicates.

The number of positive and false duplicates returned from each de-duplication option varied greatly (Table 2). The Ovid multifile search alone resulted in 1,178 citations. The comparison to the gold standard for Ovid MEDLINE and Ovid Embase revealed that simply de-duplicating in Ovid resulted in 91 false negatives but no false positives.

As mentioned above, we developed a second gold standard set for the results retrieved from Ovid MEDLINE, Ovid Embase, and CINAHL. The de-duplicated datasets from Ovid multifile and CINAHL non-MEDLINE searches, RefWorks, EndNote, and Mendeley were compared against this gold standard set. Combining the search results from the Ovid multifile search and CINAHL non-MEDLINE search options increased not only the number of false negatives by 3, but also the number of false positives by 40. De-duplicating in RefWorks resulted in 94 false negatives and 3 false positives. EndNote resulted in 258 false negatives and 6 false positives. De-duplicating with Mendeley resulted in 36 false negatives and 4 false positives.

DISCUSSION

Our primary research question was to compare the effectiveness of various de-duplication options. We were particularly interested in verifying whether using the various de-duplication options resulted in false positives (duplicates that should not have been deleted). Similar to Jiang et al., we believe false positives are more detrimental than false negatives because systematic reviewers want to maintain the highest possible recall in retrieval [7]. As running the Ovid multifile search command alone did not result in any false positives, we recommend using this option to further refine the search results before exporting to a citation manager. The limitation of this approach is that it only works if users subscribe to both MEDLINE and Embase through Ovid. PubMed users are not able to use this method.

Running the non-MEDLINE command in CINAHL, on the other hand, was the least effective method of de-duplication as it resulted in forty false positives, which was the highest number amongst all

of the options. We found that using the non-MEDLINE option in CINAHL reduced the benefit of searching multiple databases. Multi-database searching is necessary because different articles are indexed differently in different databases, so there may be articles retrieved from CINAHL that are indexed in MEDLINE but are not retrieved by the MEDLINE search. The danger of the non-MEDLINE command is that it deletes these records, reducing some of the benefit of the multi-database search.

Beyond the desire to minimize false positives, there is as yet no definitive consensus regarding how best to find and delete duplicates, although the prevalence and potential impact of duplicates remains a critical issue for those undertaking systematic reviews [8]. In 2014, Bramer et al. published a study testing the efficacy of de-duplicating with various reference managers, such as RefWorks, EndNote, Mendeley, and more [9]. According to the authors, de-duplicating exact citations in RefWorks performed the worst and de-duplicating with their proposed algorithm, named the Bramer method, yielded the best results in terms of accuracy and speed [9]. Because Bramer et al. did not distinguish the differences between false negatives and false positives, we were unable to directly compare their results to the results of our study.

A 2013 study by Qi et al. revealed that relying solely on the auto-searching feature of reference management software, such as EndNote, is inadequate when identifying duplicates for a systematic review [8].

Most recently, Rathbone et al. published a study comparing the Systematic Review Assistant-Deduplication Module (SRA-DM), a newly developed citation-screening program, against EndNote [10]. By demonstrating the superiority of the SRA-DM method, Rathbone et al.'s study also exposed the limited performance of de-duplication features in reference management software [10]. In our study, RefWorks produced the smallest number of false positives out of the citation management software that we used.

De-duplicating with Mendeley resulted in the smallest number of false negatives (citations that should have been deleted). Most notably, EndNote was the least effective citation management tool, with the highest number of false positives and false negatives. The results of our study not only confirm Qi et al.'s [8] and Rathbone et al.'s [10] findings that the automatic de-duplicating option in EndNote is inadequate and must be supplemented by hand-searching, but the results also reveal that using this option may lead to losses of articles that should not be deleted.

These data suggest that researchers will have to individually determine their own thresholds of acceptability for false positives. If none are acceptable, none of the citation management de-duplication options can be used. If the researcher is confident that all key articles would be found by hand-searching and deems a relatively low percentage of false negatives and positives to be acceptable (Table 3, online only), we recommend Mendeley as the most effective tool. Effort should be made to individually investigate all of the close duplicates in RefWorks and Mendeley to check for false positives. In addition, the results from any de-duplication technique should always be manually reviewed to check for remaining duplicates. Using formulas in Excel, such as highlighting duplicates, can be a useful tool to speed up this process.

Even with these preliminary recommendations, we must emphasize that de-duplication of results is complex. Examples of some technical issues causing difficulties in identifying duplicates automatically while creating the gold standard datasets are:

- differences in journal names (e.g., "and" instead of "&")
- punctuation (e.g., some titles are exported with a period at the end, others are not)
- translation differences of non-English article titles
- author information or order of author names

These issues are often the result of unintentional human error that occurs during the processing of individual records, and eliminating them proves challenging. Nevertheless, as commercial service providers, database administrators need to be more vigilant. Elmagarmid et al.'s article provides an extensive list of duplicate detection algorithms and metrics that can be used to clean up databases [11].

Limitations

This study does have limitations. Only two "gold standards" were used, and results may vary with other search topics. We were not able to explore the de-duplication options of other reference management software such as Zotero and Reference Manager. Future research may involve expanding the selection of reference management software.

There are many other directions that future research on this topic could take as well. For example, researchers could investigate the effectiveness of combining de-duplication codes (e.g., ..dedup) and options that can be used in bibliographic databases and refining those results with de-duplication features that various reference management software packages offer. Researchers could also test non-MEDLINE and non-MEDLINE journals commands in Embase to determine if these codes are more effective than the non-MEDLINE command in CINAHL. To foster further advancement in this field, more participation in research by librarians and information specialists is encouraged.

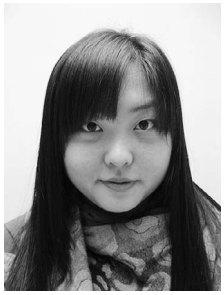
ACKNOWLEDGMENTS

The authors thank the authors of the systematic review used in this study, in particular Holly Wong, Susan E. Powelson, AHIP, Dr. William A. Ghali, and Dr. John M. Conly.

REFERENCES

1. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med. 2009 Aug 18;151(4):264–9. DOI: http://dx.doi.org/10.7326/0003-4819-151-4-200908180-00135.
2. Higgins J, Green S, eds. Cochrane handbook for systematic reviews of interventions [Internet]. Version 5.1.0. Cochrane Collaboration; 2011 [Mar 2011; cited 19 Feb 2015]. <http://www.cochrane-handbook.org>.
3. McKibbon A, Wilczynski N. PDQ evidence-based principles and practice. Shelton, CT: BC Decker/PMPH; 2009.
4. Johnson C. Repetitive, duplicate, and redundant publications: a review for authors and readers. J Manip Physiol Ther. 2006 Sep;29(7):505–9. DOI: http://dx.doi.org/10.1016/j.jmpt.2006.07.001.
5. Kassirer JP, Angell M. Redundant publication: a reminder. N Engl J Med. 1995 Aug 17;333(7):449–50. DOI: http://dx.doi.org/10.1056/NEJM199508173330709.
6. Wong H, Eso K, Ip A, Jones J, Santana M, Kwon Y, Powelson SE, De Grood J, Geransar R, Joffe AM, Taylor G, Missaghi B, Pearce C, Ghali WA, Conly JM. Unpublished data; 2015.
7. Jiang Y, Lin C, Meng W, Yu C, Cohen AM, Smalheiser NR. Rule-based deduplication of article records from bibliographic databases. Database (Oxford). 2014 Jan; 2014:bat086. DOI: http://dx.doi.org/10.1093/database/bat086.
8. Qi X, Yang M, Ren W, Jia J, Wang J, Han G, Fan D. Find duplicates among the PubMed, Embase, and Cochrane Library databases in systematic review. PLOS One.2013; 8(8):e71838. DOI: http://dx.doi.org/10.1371/journal.pone.0071838.
9. Bramer W, Holland L, Mollema J, Hannon T, Bekhuis, T. Removing duplicates in retrieval sets from electronic databases [Internet]. 2014 [cited 19 Feb 2015]. PowerPoint presentation <http://www.iss.it/binary/eahi/cont/57_Bramer_Wichor_slides_EAHIL_2014.pdf>.
10. 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: http://dx.doi.org/10.1186/2046-4053-4-6.
11. Elmagarmid AK, Ipeirotis PG, Verykios VS. Duplicate record detection: a survey. Knowl Data Eng IEEE Trans. 2007;19(1):1–16. DOI: http://dx.doi.org/10.1109/TKDE.2007.250581.

AUTHORS' AFFILIATIONS



Yoojin Kwon, MLIS, ykwon@torontopubliclibrary.ca, Librarian, Toronto Public Library, 35 Fairview Mall Drive, Toronto, ON M2J 4S4, Canada; **Michelle Lemieux, MLIS,** michelle.lemieux@altalink.ca, Regulatory Coordinator, AltaLink, 2611 Third Avenue Southeast, Calgary, AB, T2A 7W7, Canada; **Jill McTavish, PhD, MLIS,** Jill.McTavish@lhsc.on.ca, Clinical Librarian, Health Sciences Library, London Regional Cancer Program (LRCP), 790 Commissioners Road East, ON, N6A 4L6, Canada; **Nadine Wathen, PhD,** nwathen@uwo.ca, Associate Professor and Faculty Scholar, Faculty of Information & Media Studies, University of Western Ontario, North Campus Building, Room 240, London, ON, N6A 5B7, Canada

Received March 2015; accepted June 2015