ISSN 2222-0682 (online)

World Journal of *Methodology*

World J Methodol 2023 December 20; 13(5): 373-509

Published by Baishideng Publishing Group Inc

World Journal of Methodology

Contents

Quarterly Volume 13 Number 5 December 20, 2023

EDITORIAL

373 Challenges and limitations of synthetic minority oversampling techniques in machine learning

Alkhawaldeh IM, Albalkhi I, Naswhan AJ

379 Current protocol to achieve dental movement acceleration and pain control with Photo-biomodulation

Dominguez A

384 New evidence-based practice: Artificial intelligence as a barrier breaker

Ferreira RM

OPINION REVIEW

390 Evidence-based literature review: De-duplication a cornerstone for quality

Hammer B, Virgili E, Bilotta F

REVIEW

399 Crohn's disease and clinical management today: How it does?

da Silva Júnior RT, Apolonio JS, de Souza Nascimento JO, da Costa BT, Malheiro LH, Silva Luz M, de Carvalho LS, da Silva Santos C, Freire de Melo F

MINIREVIEWS

414 Using national census data to facilitate healthcare research

Colwill M, Poullis A

419 Machine learning and deep neural network-based learning in osteoarthritis knee

Ratna HVK, Jeyaraman M, Jeyaraman N, Nallakumarasamy A, Sharma S, Khanna M, Gupta A

426 Synoptic review on existing and potential sources for bias in dental research methodology with methods on their prevention and remedies

Agrawal AA, Prakash N, Almagbol M, Alobaid M, Alqarni A, Altamni H

ORIGINAL ARTICLE

Retrospective Study

439 Assessing the readability of online information about jones fracture

Al-Kharouf KFK, Khan FI, Robertson GA

446 Impact of COVID-19 lockdown on hospital admissions for epistaxis in Germany

Hoenle A, Wagner M, Lorenz S, Steinhart H

456 Effect of vaccination status on CORADS and computed tomography severity score in hospitalized COVID-19 patients: A retrospective study

Binay UD, Karavaş E, Karakeçili F, Barkay O, Aydin S, Şenbil DC

Observational Study

466 Study on good clinical practices among researchers in a tertiary healthcare institute in India

Harshita H, Panda PK

475 Inflammatory bowel disease among first generation immigrants in Israel: A nationwide epi-Israeli Inflammatory Bowel Disease Research Nucleus study

Stulman M, Focht G, Loewenberg Weisband Y, Greenfeld S, Ben Tov A, Ledderman N, Matz E, Paltiel O, Odes S, Dotan I, Benchimol EI, Turner D

Basic Study

484 Sequential extraction of RNA, DNA and protein from cultured cells of the same group

Cui YY

492 Urine exosome mRNA-based test for monitoring kidney allograft rejection: Effects of sample transportation and storage, and interference substances

McFaul M, Ventura C, Evans S, Dundar H, Rumpler MJ, McCloskey C, Lowe D, Vlassov AV

CASE REPORT

502 Successful hip revision surgery following refracture of a modern femoral stem using a cortical window osteotomy technique: A case report and review of literature

Lucero CM, Luco JB, Garcia-Mansilla A, Slullitel PA, Zanotti G, Comba F, Buttaro MA

World Journal of Methodology

Quarterly Volume 13 Number 5 December 20, 2023

Contents

ABOUT COVER

Editorial Board Member of *World Journal of Methodology*, Rodrigo Valenzuela, PhD, Associated Professor, Department of Nutrition, Faculty of Medicine, University of Chile, Independence Av. 1027, Santiago 8380000, Chile. rvalenzuelab@med.uchile.cl

AIMS AND SCOPE

The primary aim of *World Journal of Methodology* (*WJM*, *World J Methodol*) is to provide scholars and readers from various fields of methodology with a platform to publish high-quality basic and clinical research articles and communicate their research findings online.

WJM mainly publishes articles reporting research results obtained in the field of methodology and covering a wide range of topics including breath tests, cardiac imaging techniques, clinical laboratory techniques, diagnostic self-evaluation, cardiovascular diagnostic techniques, digestive system diagnostic techniques, endocrine diagnostic techniques, neurological diagnostic techniques, obstetrical and gynecological diagnostic techniques, ophthalmological diagnostic techniques, otological diagnostic techniques, radioisotope diagnostic techniques, respiratory system diagnostic techniques, surgical diagnostic techniques, *etc*.

INDEXING/ABSTRACTING

The *WJM* is now abstracted and indexed in PubMed, PubMed Central, Reference Citation Analysis, China Science and Technology Journal Database, and Superstar Journals Database.

RESPONSIBLE EDITORS FOR THIS ISSUE

NAME OF JOURNAL
World Journal of Methodology

ISSN
ISSN 2222-0682 (online)

LAUNCH DATE
September 26, 2011

FREQUENCY
Quarterly

EDITORS-IN-CHIEF
Timotius Ivan Hariyanto

POLICY OF CO-AUTHORS

EDITORIAL BOARD MEMBERS
https://www.wjgnet.com/2222-0682/editorialboard.htm

PUBLICATION DATE
December 20, 2023

INSTRUCTIONS TO AUTHORS
https://www.wjgnet.com/bpg/gerinfo/204

GUIDELINES FOR ETHICS DOCUMENTS
https://www.wjgnet.com/bpg/GerInfo/287

GUIDELINES FOR NON-NATIVE SPEAKERS OF ENGLISH
https://www.wjgnet.com/bpg/gerinfo/240

PUBLICATION ETHICS
https://www.wjgnet.com/bpg/GerInfo/288

PUBLICATION MISCONDUCT
https://www.wjgnet.com/bpg/gerinfo/208
https://www.wjgnet.com/bpg/GerInfo/310

ARTICLE PROCESSING CHARGE
https://www.wjgnet.com/bpg/gerinfo/242

STEPS FOR SUBMITTING MANUSCRIPTS
https://www.wjgnet.com/bpg/GerInfo/239

ONLINE SUBMISSION
https://www.f6publishing.com

World Journal of Methodology

Submit a Manuscript: https://www.f6publishing.com

World J Methodol 2023 December 20; 13(5): 390-398

DOI: 10.5662/wjm.v13.i5.390

ISSN 2222-0682 (online)

OPINION REVIEW

Evidence-based literature review: De-duplication a cornerstone for quality

Barbara Hammer, Elettra Virgili, Federico Bilotta

Specialty type: Medical laboratory technology

Provenance and peer review: Invited article; Externally peer reviewed.

Peer-review model: Single blind

Peer-review report's scientific quality classification

Grade A (Excellent): 0
Grade B (Very good): B
Grade C (Good): 0
Grade D (Fair): 0
Grade E (Poor): 0

P-Reviewer: Morya AK, India

Received: October 10, 2023
Peer-review started: October 10, 2023
First decision: October 24, 2023
Revised: November 3, 2023
Accepted: November 29, 2023
Article in press: November 29, 2023
Published online: December 20, 2023

Barbara Hammer, Librarian at Medical Library, University of Bergen, Bergen 5020, Norway

Elettra Virgili, Federico Bilotta, Anesthesiology, Critical Care and Pain Medicine, University of Rome "La Sapienza", Rome 00166, Italy

Corresponding author: Federico Bilotta, MD, PhD, Professor, Anesthesiology, Critical Care and Pain Medicine, University of Rome "La Sapienza", viale del Policlinico, 155, Rome 00166, Italy. federico.bilotta@uniroma1.it

Abstract

Evidence-based literature reviews play a vital role in contemporary research, facilitating the synthesis of knowledge from multiple sources to inform decision-making and scientific advancements. Within this framework, de-duplication emerges as a part of the process for ensuring the integrity and reliability of evidence extraction. This opinion review delves into the evolution of de-duplication, highlights its importance in evidence synthesis, explores various de-duplication methods, discusses evolving technologies, and proposes best practices. By addressing ethical considerations this paper emphasizes the significance of de-duplication as a cornerstone for quality in evidence-based literature reviews.

Key Words: Duplicate publications as topic; Databases; Bibliographic; Artificial intelligence; Systematic reviews as topic; Review literature as topic; De-duplication; Duplicate references; Reference management software

Core Tip: Effective de-duplication is crucial for maintaining the quality and credibility of systematic reviews. It ensures data accuracy, eliminates bias, reduces workload, and enhances trust in findings. However, challenges such as variability in data, database indexing, and resource constraints exist. Best practices include clear documentation, the use of reference management software, manual review when necessary, handling multiple versions of the same paper, addressing non-journal sources, and ethical considerations. Advancements like Deduklick and Automated Systematic Search Deduplicator offer promise for more accurate and efficient de-duplication methods. De-duplication is a fundamental step in evidence synthesis, contributing to transparent and reproducible research in systematic reviews.

Citation: Hammer B, Virgili E, Bilotta F. Evidence-based literature review: De-duplication a cornerstone for quality. *World J Methodol* 2023; 13(5): 390-398
URL: https://www.wjgnet.com/2222-0682/full/v13/i5/390.htm
DOI: https://dx.doi.org/10.5662/wjm.v13.i5.390

INTRODUCTION

Evidence-based literature reviews are essential for informed decision-making in research and practice. However, without proper de-duplication, duplicated records may skew findings, leading to biased conclusions. This opinion review aims to shed light on the importance of de-duplication in evidence synthesis and its evolution over time and to give a big picture of what's available as a solution so research teams can make an informative decision on what's best for the project when it comes to de-duplication.

What is de-duplication?

In the world of computing, database de-duplication refers to the technique of ensuring that specific information is only stated once. One way to find information that is consistent across multiple sources (such as data files, books, websites, and databases) is through record linkage (RL). Data matching, RL, data linkage, entity resolution, and many other terms are focused on finding such records and eliminating duplicates. "RL is necessary when joining different data sets based on entities that may or may not share a common identifier (*e.g.*, database key, URI, National identification number), which may be due to differences in record shape, storage location, or curator style or preference"[1]. The term de-duplication, in medical scientific writing, refers to the process of identifying and removing duplicate citations from the search results retrieved from various databases. "Record de-duplication is of great advantage for de-duplicating citation in bibliographic databases"[2]. However, the identification of duplicated citations is not a trivial task. "Records are usually not identical, because they may come from different databases and may differ in the treatment of authors' names of journal titles, indexing, and special field"[3]. Duplicated citations are the result of the standard in evidence synthesis studies like systematic reviews which require comprehensive searching in multiple databases to identify eligible studies. Duplicates can arise due to multiple factors, such as the same study being indexed in multiple databases and because "certain types of information are recorded differently (and inconsistently) in the different databases"[4].

As stated by the research team from the Pain Research, Nuffield Department of Anaesthetics, Churchill Hospital in Oxford: "if duplicate records are not removed effectively, reviewers can waste time screening the same records for inclusion and run the risk of accidentally including same paper more than once in their meta-analyses, leading to inaccurate conclusions"[5]. Hence removing duplicate citations is an important and necessary step between searching and screening in a process of the systematic review.

In practice, de-duplication is available *via* search platforms, reference management software (Table 1), and screening assistance tools (Table 2). However, automation does not entirely solve de-duplication issues. The process typically involves exporting search results into the library in one of the research management tools, merging data sets and identifying duplicated citations. Duplicates are identified by comparing various bibliographic elements such as titles, authors, publication dates, journal names, *etc.* Once potential duplicates are identified, the research team reviews these records to confirm if they are indeed duplicates. Confirmed duplicates are removed from the data set, ensuring that only unique study is kept and counted only once in the systematic review. Then the de-duplicated data set forms the basis for subsequent stages of the systematic review: study selection, data extraction, and data synthesis.

EVALUATION OF THE DE-DUPLICATION - METHODS, TECHNIQUES, AND TOOLS

De-duplication can be traced back to various stages in the development of information management and technology. Over time, de-duplication methods have evolved from manual to sophisticated automated techniques. However, to this day effective de-duplication methods may involve a combination of automated and manual approaches to ensure accurate and reliable results, as the race to create the ultimate tool continues. "In the early days of bibliographic record-keeping manual cataloging was the only process of creating metadata representing information resources, such as books, sound recordings, moving images, *etc.* Cataloging provided information such as author's names, titles, and subject terms that describe resources, typically through the creation of bibliographic records"[6]. Librarians and researchers manually

Table 1 Reference management software

Paid	Free
EndNote	Mendeley
RefWorks	Zotero
	Bib TeX

Table 2 Systematic review tools that offer de-duplication

Paid	Free
Covidence	Rayyan
DistillerSR	Systematic Review Assistant-De-duplication Module
Deduklick	Automated Systematic Search Deduplicator

reviewed catalogue cards or printed bibliographies to identify and eliminate duplicate entries. "The introduction of the term database coincided with the availability of direct-access storage from the mid-1960s onwards. As computers grew in speed and capability, several general-purpose database systems emerged; by the mid-1960s a number of such systems had come into commercial use"[7]. "Since the 1970s metadata were in machine-readable form and were indexed by information retrieval tools, such as bibliographic databases or search engines"[6].

With the rise of electronic databases, de-duplication has become more complex. Databases allowed for the storage and retrieval of vast amounts of bibliographic information, leading to an increased need for automated methods in de-duplication. There are several existing de-duplication tools, methods, and techniques that have been developed to address the challenges of identifying and eliminating duplicates from datasets. "Reference management software has been a useful tool for researchers since the 1980s"[8]. Within a brief timeframe, a marketplace was established, and commercial products were produced. The development of *e.g.*, EndNote[9], Zotero[10], or Mendeley[11], *etc.* provided researchers with tools to manage and de-duplicate their collections of references. This type of tool introduced automated (default) de-duplication features.

The introduction of Digital Object Identifiers (DOIs) "in the late 1990s, and implementation in the early 2000s"[12], has greatly facilitated de-duplication. DOIs provide a standardized and unique identifier for each publication, making it easier to track and manage duplicates. Screening assistance tools like, *e.g.*, Covidence[13] and Rayyan[14] were developed specifically for systematic reviews. These tools integrated automated de-duplication, collaboration features, and support for the review process. "Modern data matching algorithms utilize advanced techniques, including tokenization, stemming, and phonetic algorithms, to handle variations in text data and improve the accuracy of matching.

The following terms explain the various types of de-duplication processes: (1) Exact match de-duplication: This method examines precise matches in key fields, such as unique identifiers or customer IDs. If the same information is shown on multiple records, these duplicates are removed; (2) Fuzzy match de-duplication: Fuzzy de-duplication techniques use algorithms to determine the similarity between records, even if they do not have exact matches in key fields, allowing for the recognition of duplicates with slight differences or misspellings; and (3) Rule-based de-duplication: Rule-based de-duplication involves defining specific rules or criteria to identify duplicates. These rules can be based on data patterns, business logic, or specific requirements"[15].

Practical solutions for effective de-duplication constantly evolve. In 2013 research team from Fourth Military Medical University in China described a pragmatic strategy of combining automated and manual searching duplicates in a systematic review[16]. This paper evaluates the extensiveness and characteristics of duplicates in the PubMed, EMBASE, and Cochrane Library databases. Identifies two types of duplicates: Type-I (duplicates among different databases) and type-II (duplicate publications in different journals/issues). Results showed that most type-I duplicates are identified by the auto-searching method, while nearly all type-II duplicates are identified by the hand-searching method. The hand-searching approach has a substantially greater incidence of incorrect items in type-I duplicates, most of which come from the EMBASE database. The authors recommend employing a combined strategy of auto-and-hand-searching methods to find duplicates in the systematic review due to the insufficiency of a single strategy.

In 2015, Canadian researchers explored and compared the effectiveness of various de-duplication features, specifically in the Ovid and EBSCO database platforms and three selected reference management software packages: RefWorks, EndNote, and Mendeley[17]. The authors recorded the time taken to de-duplicate each option, the number of false positives, and the false negatives, and in conclusion, recommended different de-duplication options based on the skill of the searcher and the reason for de-duplication. Overall, the results of the study highlight the variation in time and effectiveness of different de-duplication options, providing insights for researchers to choose the most suitable option based on their needs and expertise. Same year research team from Bond University in Australia developed a de-duplication program - The Systematic Review Assistant-De-duplication Module (SRA-DM) to improve the effectiveness of duplicate detection[18]. The paper presents the evaluation of the SRA-DM against EndNote's default de-duplication process, comparing their sensitivity and specificity in detecting duplicates. The goal of the study was to determine the reliability and effectiveness of the SRA-DM in removing duplicate records. In conclusion, SRA-DM demonstrated superior

sensitivity (84%) and specificity (100%) compared to EndNote's default de-duplication process, resulting in a 42.86% increase in the number of duplicate records detected. The paper acknowledged that no software can currently detect all duplicate records, and there are limitations to the SRA-DM, such as undetected duplicates due to discrepancies in data and extraneous information inserted into the title field.

In 2016, an international research team led by Erasmus MC-Erasmus University Medical Centre in Rotterdam developed a de-duplication method (colloquially referred to as the Bramer method) for de-duplicating database search results in EndNote, a popular reference manager, which is used by information professionals conducting exhaustive searches for systematic reviews[19]. The authors highlight the limitations of relying on unique identifiers like DOIs and PMIDs for identifying duplicates and propose using pagination as an alternative. They discuss the variations in page number formats used in different databases and provide a method for adapting the page number format of references to facilitate de-duplication. The paper addresses the challenges of existing de-duplication methods, which are time-consuming or impractical, and compares different software programs. The authors provide detailed instructions for customizing EndNote settings, creating export files with expanded page numbers, and installing filters for importing modified files. Overall, the paper contributes a practical and efficient method for de-duplicating database search results in EndNote, addressing the limitations and challenges of existing methods. This method is still very popular even though it was introduced in 2016.

In 2019, two researchers from the university library at the Vrije University in Amsterdam created AMSTERDAM EFFICIENT DE-DUPLICATION (AED) METHOD. The paper describes the authors' method of de-duplication, which provides a systematic approach to de-duplicating articles and claims to be 100% reliable[20]. The AED method explains per database/host what steps are needed to successfully de-duplicate data sets. This multi-step approach for efficient de-duplication includes collecting accession numbers during the initial search which is useful for an update search and then followed by manual assessment. If the data set is large authors advise following up with the Bramer method.

In 2021, Canadian researchers evaluated the accuracy and efficiency of commonly used electronic methods for flagging and removing duplicate references in systematic reviews[21]. Testing included the default settings (using the default algorithm of each program) in Ovid multifile search, EndNote desktop, Mendeley, Zotero, Covidence, and Rayyan. A benchmark set of unique, de-duplicated references was created through manual abstraction, and the performance of different de-duplication methods was compared against this benchmark set. The study identifies Ovid, Covidence, and Rayyan as the most accurate methods for identifying duplicate references, with Ovid and Covidence having high specificity and Rayyan demonstrating high sensitivity. The paper highlights the strengths and weaknesses of commonly used de-duplication methods and provides strategies for improving their performance to avoid unintentionally removing eligible studies and introducing bias into systematic reviews. The limitation of this paper is the fact that it does not provide specific details about the number of false-negative and false-positive duplicate references for each method or the overall accuracy, sensitivity, and specificity values. Still, the findings of the study are important for researchers in selecting database platforms and supporting software programs for conducting systematic reviews, highlighting those factors such as availability, ease of use, functionality, and capability must be taken into consideration.

In 2022, researchers were introduced to the most advanced de-duplication to date. The Swiss research team from the University of Bren has developed an automated, artificial intelligence-based algorithm named "Deduklick" which combines natural language processing algorithms with a set of rules created by expert information specialists[22]. This automated de-duplication uses a multistep algorithm of data normalization, calculates a similarity score, and identifies unique and duplicate references based on metadata fields, such as title, authors, journal, DOI, year, issue, volume, and page number range. Authors claim that the algorithm significantly reduced the time spent on analysis, simplifying the systematic review process. The performance was comparable to expert information specialists while preserving high metadata quality. The algorithm's transparent and explainable decision process, along with its reproducibility and adherence to Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) standards[23], makes it a reliable tool for duplicate removal. Although this sounds groundbreaking, this paper is not free from limitations *e.g.*, it does not address potential biases or limitations in the algorithm's decision-making process, such as the impact of different metadata fields on duplicate detection or does not discuss potential limitations or challenges in implementing Deduklick in real-world systematic review processes. Also, the evaluation of the performance was limited to eight existing datasets, and it is unclear if these datasets represent the full range of systematic review scenarios.

In 2023, research team from the University of Edinburgh developed Automated Systematic Search Deduplicator (ASySD) an automated tool to conduct the de-duplication of systematic searches in biomedical databases[24]. In this paper, the authors compared ASySD with other existing tools, such as EndNote's default de-duplication option and the SRA-DM. As a result, ASySD outperformed the alternative methods, correctly removing > 95% of duplicate citations across five biomedical datasets, while removing a few citations incorrectly (specificity > 0.999). The paper's limitations are - the tool was only developed using preclinical systematic review datasets and its performance in other review areas has not been fully evaluated. The accuracy of ASySD depends on the quantity and quality of citation information, and it may not work well with older searches or citations lacking certain bibliographic information. ASySD may mistakenly remove some citations, which is a concern for smaller reviews where each relevant paper is important. Additionally, the memory requirements for larger datasets exceed what is possible on shinyapps.io, so users may need to run the Shiny application locally in R Studio, which can be challenging for non-R-proficient individuals.

IMPORTANCE OF DE-DUPLICATION IN EVIDENCE SYNTHESIS

De-duplication serves as an important step in evidence-based literature reviews hence research teams must be aware of

the two types of de-duplication in medical scientific writing[25]. As mentioned earlier the first type exists on a database level and is a result of the event when a single manuscript was concurrently entered into two or more databases. Usually detected *via* automatic de-duplication. The second type exists between various journals when a single manuscript/study was released in several publications/issues/editions. This category is often referred to as study de-duplication "as it aims to identify two distinct reports of the same study. Study deduplication, although the rarer of the two is only usually detected after data have been extracted from both papers, after the authors have been contacted, or sometimes not at all" [26]. Usually detected *via* screening in full text. Removing duplicate records is essential in maintaining data accuracy and integrity, as they can introduce inaccuracies and redundant data extraction and analysis efforts. De-duplication streamlines the process by focusing on unique evidence, minimizing confusion and misinterpretation in systematic reviews. It eliminates bias and overestimation as duplicates can bias the results of evidence synthesis by inflating the apparent number of studies available for analysis. "Using a primary study results multiple times in the same analysis overstates its sample size and number of events, falsely leading to greater precision in the analysis"[27]. It enhances the quality of synthesis as evidence synthesis aims to provide a comprehensive and accurate overview of the available evidence. "If the same study has more than one report - possibly with different author lists, different titles, and in different journals - both papers should often be cited, but they should only be included in the meta-analysis as one trial" [26,28]. De-duplication ensures that the synthesis is based only on a unique and high-quality set of studies. Reduce workload and improve efficiency as removing duplicates reduces the workload for reviewers and analysts, allowing them to focus on analysing unique studies. This enhances the efficiency of the evidence synthesis process. Provide transparent and reproducible research as de-duplication is an integral part of transparent and reproducible research. Documenting the process ensures that others can replicate the de-duplication and validation steps, enhancing the credibility of the review. Align with publication standards and guidelines as many publication guidelines, including the PRISMA, emphasize the importance of de-duplication to maintain the quality and credibility of systematic reviews. Most importantly enhances trust in findings as de-duplication increases the trustworthiness of the systematic review findings by demonstrating a rigorous and transparent approach to handling data.

WHAT ARE THE CHALLENGES AND LIMITATIONS OF DE-DUPLICATION IN SYSTEMATIC REVIEWS?

De-duplication in systematic reviews, while necessary, is not without challenges and limitations which were touched on earlier. It is important to be aware of these potential issues to effectively address them and ensure the accuracy and reliability of the review process. Conducting manual reviews to confirm duplicates is time-consuming, especially when titles and abstracts do not provide enough information for differentiation. Managing and processing large complex data sets manually can be time-consuming, and error-prone, automated de-duplication seems to be the best option however deciding on either manual or automated de-duplication requires careful consideration.

Variability in data contributes to the issue - variations in study titles, author names, and other bibliographic elements can complicate the de-duplication process as it makes it difficult to accurately identify duplicates requiring careful consideration of matching criteria.

Differences in database indexing - different databases use varied indexing and citation formats which leads to inconsistent or incomplete data *e.g.*, page numbers, which can affect how duplicates are identified, making it harder to determine if two records are indeed duplicates. Furthermore, as a result of those discrepancies automated de-duplication tools may produce "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)"[17]. Hence researchers need to account for these variations during the de-duplication process and often manual review is essential to confirm results.

Cross-language studies - dealing with studies published in different languages introduces challenges due to variations in titles, authors, and other bibliographic details. Non-journal sources - systematic reviews may include various types of sources beyond journal articles *e.g.*, reports, theses, and conference papers. These sources may have different indexing and citation formats, making de-duplication more complex.

Multiple versions of the same studies also contribute to the problem. Studies may be published in different versions (*e.g.*, conference abstracts, and full-text articles) or have been published in multiple journals. Deciding whether these are duplicates or unique records requires careful assessment as "there is currently no standard methodological approach to deal with overlap in primary studies across reviews"[27].

Risk of exclusion - overly aggressive de-duplication can lead to the inadvertent exclusion of potentially relevant studies, affecting the comprehensiveness of the review. Technological limitations - automated tools may not be able to handle certain complexities, such as very similar studies with nuanced differences *e.g.*, differences in journal names "and" instead of "&" or author information or order of authors names[17].

Resource constraints - limited access to paid-for automated tools, lack of personnel, or time can impact the thoroughness of the de-duplication process.

Striking a balance between efficiency and accuracy is essential to overcome these limitations. Navigating these challenges requires a combination of methodological rigor, technological tools, collaboration among the review team, and transparent reporting of the de-duplication process and outcomes.

THE ROLE OF PROSPECTIVE REGISTRATION OF SYSTEMATIC REVIEWS AND META-ANALYSIS AND HOW THIS HELPS IN DE-DUPLICATION?

Prospective registration involves registering systematic reviews and meta-analyses in publicly accessible databases before starting the research process. This practice has gained prominence in recent years, primarily due to its significant impact on de-duplication efforts. In a 2022 paper by a German team from Brandenburg Medical School (Theodor Fontane)[29], the authors stated that prospective registration of systematic reviews aims to reduce bias in research conduct and reporting, increase transparency, and prevent unintended duplication, thereby reducing research waste. There are several options available for prospective registration, including PROSPERO, the Registry of Systematic Reviews/Meta-Analyses in Research Registry, INPLASY, the Open Science Framework Registries, and protocols.io. These registries provide search functions to help authors avoid duplicate reviews.

Prospective registration discourages the submission of the same systematic review or meta-analysis to multiple journals, as researchers and publishers can easily identify prior registrations. Hence reduces the chances of duplicate publications, a common issue in medical literature, which can subsequently lead to de-duplication problems. Registered systematic reviews and meta-analyses are required to provide a detailed protocol outlining their research objectives, methodologies, and inclusion criteria. This transparency helps researchers identify potentially duplicate records, even before data collection begins. Prospective registration fosters collaboration by allowing other researchers to see that other reviews are ongoing or coming up in relation to their own field. But also fosters group work and discourages the chances of having redundant reviews at the same time.

THE ROLE OF REFERENCE CITATION ANALYSIS FOR PROPER CITATION AND DE-DUPLICATION

Another important tool to improve de-duplication in medical databases is reference citation analysis and this goes hand in hand with prospective registration. "Use of the unique registration number may be useful in helping track subsequent use or citation of the review to monitor its impact"[30]. It involves a meticulous examination of the references cited in articles, and it plays a critical role as *via* reference citation analysis, researchers can identify secondary publications that stem from the same primary research, such as conference abstracts, journal articles, and systematic reviews. This is crucial for de-duplication, as it helps consolidate related information into a single reference. Citation analysis also aids in ensuring that the primary sources are correctly attributed and cited in systematic reviews and meta-analyses. But also, can reveal citation errors, discrepancies, or inconsistencies in systematic reviews and meta-analyses. Identifying and rectifying these issues contribute to the overall quality of the research synthesis. This helps maintain accuracy and integrity in the research synthesis process.

BEST PRACTICES FOR DE-DUPLICATION IN LITERATURE REVIEWS

De-duplicating search results and studies effectively during the systematic review process is essential. A comprehensive understanding of the data set's characteristics and proper validation of de-duplication outcomes are also critical. Transparent documentation of de-duplication procedures and reproducibility of results are important. Yet there are no standardized guidelines for all aspects of de-duplication, leading to variations in practices and interpretations as shown in this opinion literature review.

Following best practices helps maintain the quality of the review and the credibility of its findings. The practice proposed in this opinion review reflects the personal approach of the librarian but can be applied broadly to all researchers working on a systematic review to achieve reliable and reproducible results.

Document the process - clearly document your de-duplication process in the review protocol. This documentation should include the criteria for identifying duplicates, the tools/software used, and any decisions made during the process of de-duplication so it could be replicated by others. Transparency in the methodology enhances the credibility of the review.

Utilize reference management software (*e.g.*, EndNote, Zotero, Mendeley) to manage and organize search results. These tools include automated (default) de-duplication features that help identify at least the exact matches and reduce obvious duplicates. *e.g.*, in EndNote the default settings are author, year, and title. These tools can also further help identify duplicates based on predefined criteria *e.g.*, volume, issue, and pages which require deciding on a method that is best for that project *e.g.*, the Bramer method for EndNote.

Manual review - conduct a manual review of potential duplicates identified by some of the automated tools. Establish criteria for matching - define explicit criteria for matching *e.g.*, titles, authors, publication dates, and other bibliographic information to confirm whether records are indeed duplicates. Decide on a threshold for matching to avoid excluding potentially relevant studies.

Handle multiple versions of the same paper - pay attention to different versions *e.g.*, conference abstracts, and full-text articles. Decide whether to treat them as separate records from the start or duplicates based on their content (usually screening in full text will solve this problem).

Address non-journal sources - be prepared to de-duplicate various types of sources beyond journal articles, such as conference proceedings, reports, and theses. Consider their unique indexing and citation formats.

Handle updates and overlapping searches of the existing systematic review strategy. If your review involves multiple search rounds, use the de-duplication process to identify studies already included in previous rounds to avoid the common assumption that updating the search strategy is as easy as taking it from where you left off.

Resolve discrepancies - in case of discrepancies or uncertainties, consult with your review team to make informed decisions about the status of potentially duplicated records. Document decisions - document all decisions made during the de-duplication process, including the rationale for excluding or retaining records. Transparency in decision-making enhances the review's reproducibility.

Remember that while automated tools can expedite the de-duplication process, often manual review is unavoidable and still crucial for accurate identification of duplicates, especially when titles and abstracts are not sufficient for differentiation. Consistency, thoroughness, and transparency are key principles when de-duplicating studies in systematic reviews.

ETHICAL CONSIDERATIONS IN DE-DUPLICATION

While de-duplication is essential for data integrity and research quality, it's important to approach the task in a manner that respects the rights of authors and researchers, maintains data privacy, and adheres to ethical standards *e.g.*, the European Code of Conduct of the Research Integrity published by All European Academies[31]. Ethical considerations in de-duplication align with principles of reliability, honesty, respect, and accountability in responsible research and data management outlined by the above code.

With that in heart, this opinion review proposes the following considerations, transparency and documentation: (1) Systematic review protocol should provide transparent information about the de-duplication process, including the criteria used, methods applied, and decisions made. Transparent documentation helps ensure accountability and reproducibility, allowing others to understand and verify the process; (2) Preservation of data integrity: While removing duplicates is necessary, ensure that the process does not alter or compromise the integrity of the original data. Keep the original data as a reference for any future inquiries; (3) Conflict resolution: In cases of disagreements or uncertainties about the status of a record, aim for consensus within the review team to resolve conflicts ethically and responsibly; and (4) Maintaining original records: Keep a copy of the original duplicated records, even if they are removed from the final dataset. This preserves a historical record of the research process.

FUTURE DIRECTIONS AND CHALLENGES

The future of de-duplication holds exciting possibilities as technology continues to evolve. Continued advancements in machine learning, deep learning, and natural language processing will enable more accurate de-duplication. Deduklick is already on that path as its first de-duplication tool to ease the de-duplication burden. Standardization efforts to harmonize data formats, identifiers, and metadata across different sources would also simplify de-duplication processes. Collaboration across disciplines, ongoing research, and innovative solutions will even further shape the future of de-duplication.

CONCLUSION

Accurate and reliable de-duplication stands as a cornerstone for quality in evidence-based literature reviews. By addressing issues of duplicate records and data redundancies, de-duplication plays a critical role in upholding the scientific rigor, transparency, and overall quality of systematic reviews, making them more trustworthy and impactful resources for evidence-based decision-making. Although this functionality is available *via* many tools not all of them keep up with current advancements in the field of computer science and continue to see de-duplication only as one of many functions' tools were designed to perform. With Deduklick and AsySD the future of de-duplication holds promise for more accurate, efficient methods that can handle increasingly complex and diverse datasets.

FOOTNOTES

Author contributions: Hammer B contributed to the literature search; Virgili E involved in the extraction; Bilotta F participated in the project design data revision; Hammer B, Virgili E, and Bilotta F wrote the manuscript; and all authors have read and approved the final manuscript.

Conflict-of-interest statement: All the authors report no relevant conflicts of interest for this article.

Country/Territory of origin: Italy

ORCID number: Barbara Hammer 0009-0001-6361-0880; Elettra Virgili 0009-0000-9108-9570; Federico Bilotta 0000-0003-2496-6646.

S-Editor: Wang JJ
L-Editor: A
P-Editor: Xu ZH

REFERENCES

1 **Wikipedia contributors**. Deduplication. Wikipedia, The Free Encyclopaedia. [cited 27 September 2023]. Available from: https://en.wikipedia.org/w/index.php?title=Deduplication&oldid=920005448
2 **Sohail A**, Yousaf MM. A proficient cost reduction framework for de-duplication of records in data integration. *BMC Med Inform Decis Mak* 2016; **16**: 42 [PMID: 27067004 DOI: 10.1186/s12911-016-0280-9]
3 **McGowan J**, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS Peer Review of Electronic Search Strategies: 2015 Guideline Statement. *J Clin Epidemiol* 2016; **75**: 40-46 [PMID: 27005575 DOI: 10.1016/j.jclinepi.2016.01.021]
4 **Jiang Y**, Lin C, Meng W, Yu C, Cohen AM, Smalheiser NR. Rule-based deduplication of article records from bibliographic databases. *Database (Oxford)* 2014; **2014**: bat086 [PMID: 24434031 DOI: 10.1093/database/bat086]
5 **Tramèr MR**, Reynolds DJ, Moore RA, McQuay HJ. Impact of covert duplicate publication on meta-analysis: a case study. *BMJ* 1997; **315**: 635-640 [PMID: 9310564 DOI: 10.1136/bmj.315.7109.635]
6 **Wikipedia contributors**. Cataloging (library science). Wikipedia, The Free Encyclopedia. [cited 28 September 2023]. Available from: https://en.wikipedia.org/w/index.php?title=Cataloging_(library_science)&oldid=1169608000
7 Wikipedia contributors. Database. Wikipedia, The Free Encyclopedia. [cited 27 September 2023]. Available from: https://en.wikipedia.org/w/index.php?title=Database&oldid=1171956467
8 **Tramullas J**, Sánchez-Casabón AI, Garrido-Picazo P. Studies and Analysis of Reference Management Software: A Literature Review. *El profesional de la información* 2015 [DOI: 10.3145/epi.2015.sep.17]
9 Clarivate Analytics. EndNote [Internet]. [cited 20 September 2023]. Available from: https://endnote.com/?language=en
10 Zotero Groups. [cited 20 September 2023]. Available from: https://www.zotero.org/groups/
11 Mendeley. [cited 20 September 2023]. Available from: https://www.mendeley.com/
12 American University. Digital Object Identifiers and their use at American U.: DOIs. [cited 20 September 2023]. Available from: https://subjectguides.library.american.edu/DOIs
13 Convidence. The world's Systematic Review Tool. [cited 20 September 2023]. Available from: https://www.covidence.org/about-us-covidence/
14 Rayyan. [internet]. [cited 20 September 2023]. Available from: https://www.rayyan.ai/
15 Dremio. Deduplication. [cited 20 September 2023]. Available from: https://www.dremio.com/wiki/deduplication/
16 **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**: e71838 [PMID: 23977157 DOI: 10.1371/journal.pone.0071838]
17 **Kwon Y**, Lemieux M, McTavish J, Wathen N. Identifying and removing duplicate records from systematic review searches. *J Med Libr Assoc* 2015; **103**: 184-188 [PMID: 26512216 DOI: 10.3163/1536-5050.103.4.004]
18 **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**: 6 [PMID: 25588387 DOI: 10.1186/2046-4053-4-6]
19 **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**: 240-243 [PMID: 27366130 DOI: 10.3163/1536-5050.104.3.014]
20 **Otten R**, de Vries R, Schoonmade L. Amsterdam Efficient Deduplication (AED) method. *Zenodo* 2019
21 **McKeown S**, Mir ZM. Considerations for conducting systematic reviews: evaluating the performance of different methods for de-duplicating references. *Syst Rev* 2021; **10**: 38 [PMID: 33485394 DOI: 10.1186/s13643-021-01583-y]
22 **Borissov N**, Haas Q, Minder B, Kopp-Heim D, von Gernler M, Janka H, Teodoro D, Amini P. Reducing systematic review burden using Deduklick: a novel, automated, reliable, and explainable deduplication algorithm to foster medical research. *Syst Rev* 2022; **11**: 172 [PMID: 35978441 DOI: 10.1186/s13643-022-02045-9]
23 PRISMA. Welcome to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) website! [cited 20 September 2023]. Available from: http://prisma-statement.org/Default.aspx
24 **Hair K**, Bahor Z, Macleod M, Liao J, Sena ES. The Automated Systematic Search Deduplicator (ASySD): a rapid, open-source, interoperable tool to remove duplicate citations in biomedical systematic reviews. *BMC Biol* 2023; **21**: 189 [PMID: 37674179 DOI: 10.1186/s12915-023-01686-z]
25 **Qi XS**, Bai M, Yang ZP, Ren WR. Duplicates in systematic reviews: A critical, but often neglected issue. *World J Meta-Anal* 2013; **1**: 97-101 [DOI: 10.13105/wjma.v1.i3.97]
26 **Tsafnat G**, Glasziou P, Choong MK, Dunn A, Galgani F, Coiera E. Systematic review automation technologies. *Syst Rev* 2014; **3**: 74 [PMID: 25005128 DOI: 10.1186/2046-4053-3-74]
27 **Lunny C**, Pieper D, Thabet P, Kanji S. Managing overlap of primary study results across systematic reviews: practical considerations for authors of overviews of reviews. *BMC Med Res Methodol* 2021; **21**: 140 [PMID: 34233615 DOI: 10.1186/s12874-021-01269-y]
28 **Aabenhus R**, Jensen JU, Cals JW. Incorrect inclusion of individual studies and methodological flaws in systematic review and meta-analysis. *Br J Gen Pract* 2014; **64**: 221-222 [PMID: 24771816 DOI: 10.3399/bjgp14X679615]
29 **Pieper D**, Rombey T. Where to prospectively register a systematic review. *Syst Rev* 2022; **11**: 8 [PMID: 34998432 DOI: 10.1186/s13643-021-01877-1]
30 **Stewart L**, Moher D, Shekelle P. Why prospective registration of systematic reviews makes sense. *Syst Rev* 2012; **1**: 7 [PMID: 22588008 DOI:

10.1186/2046-4053-1-7]

31 **All European Academies**. European Code of Conduct for Research Integrity. [cited 20 September 2023]. Available from: https://allea.org/code-of-conduct/

Published by **Baishideng Publishing Group Inc**
7041 Koll Center Parkway, Suite 160, Pleasanton, CA 94566, USA
Telephone: +1-925-3991568
E-mail: office@baishideng.com
Help Desk: https://www.f6publishing.com/helpdesk
https://www.wjgnet.com