Esc Anna Nery 2023;27:e20220143

Validity assessment of a computational system in the identification of duplicate studies

Avaliação de validade de um sistema computacional na identificação de estudos duplicados

Evaluación de la validez de un sistema computacional en la identificación de estudios duplicados

Fernanda Martins Dias Escaldelai[1]
Leandro Escaldelai[2]
Denise Pimentel Bergamaschi[1]

1. Universidade de São Paulo. São Paulo, SP, Brasil.

2. Faculdade de Tecnologia de São Paulo. São Paulo, SP, Brasil.

Abstract

Objective: To evaluate the performance of the Systematic Review Support web-based system for the identification of duplicate records compared with similar software tools. **Methods:** A methodological study was conducted assessing the automated process of de-duplication performed by the Systematic Review Support web-based system (version 1.0) versus the EndNote X9® and Rayyan® systems, adopting hand-checking as the benchmark reference for comparisons. A set of studies on three topics related to cystic fibrosis retrieved from the Pubmed, Embase and Web of Science electronic databases was used for testing purposes. The sensitivity, specificity, accuracy and area under the ROC curve of the software systems were compared to the benchmark values for performance evaluation. **Results:** The database searches retrieved 1332 studies, of which 273 (20.5%) were true duplicates. The Systematic Review Support tool identified a larger proportion of true duplicates than the other systems tested. The sensitivity, specificity and accuracy of the Systematic Review Support tool exceeded 98%. **Conclusion and implications for practice:** The Systematic Review Support system provided a high level of sensitivity, specificity and accuracy in identifying duplicate studies, optimizing time and effort by reviewers in the health field.

Keywords: Data Accuracy; Databases, Bibliographic; Systematic Review; Sensitivity and Specificity; Software.

Resumo

Objetivo: Avaliar o desempenho do sistema *web* "Apoio à Revisão Sistemática" quanto à identificação de referências bibliográficas duplicadas, em comparação a outros programas. **Métodos:** Trata-se de uma pesquisa metodológica que avalia o processo automático de identificação de duplicatas do sistema "Apoio à Revisão Sistemática" (versão 1.0), em comparação ao *EndNote X9®* e *Rayyan®*, considerando checagem manual como referência. Foi utilizado um conjunto de estudos relacionados a três temas sobre fibrose cística recuperados das bases de dados *Pubmed, Embase* e *Web of Science*. Para avaliação de desempenho, utilizaram-se a sensibilidade, especificidade, acurácia e área sob a curva ROC para cada *software,* em comparação à referência. **Resultados:** As buscas nas bases de dados resultaram em 1332 estudos, sendo 273 (20,5%) verdadeiros duplicados. Em comparação aos dados de referência, o programa "Apoio à Revisão Sistemática" identificou maior proporção de duplicatas verdadeiras do que os demais. Os valores de sensibilidade, especificidade e acurácia do sistema "Apoio à Revisão Sistemática" apresentaram-se acima de 98%. **Conclusão e implicações para a prática:** O sistema "Apoio à Revisão Sistemática" possui alta sensibilidade, especificidade e acurácia para identificação de estudos duplicados, otimizando o tempo e o trabalho dos revisores da área da saúde.

Palavras-chave: Acurácia dos Dados; Bases de Dados Bibliográficos; Revisão Sistemática; Sensibilidade e Especificidade; Software.

Resumen

Objetivo: Evaluar el desempeño del sistema *web* "Apoyo a la Revisión Sistemática" en cuanto a la identificación de referencias duplicadas en comparación a otros programas. **Métodos:** Se trata de una investigación metodológica que evalúa el proceso automático de desduplicación del sistema *web* "Apoyo a la Revisión Sistemática" (versión 1.0), en comparación al EndNote X9® y Rayyan®, considerando la verificación manual como referencia. Fue utilizado, como ejemplo, un conjunto de estudios relacionados a tres temas sobre fibrosis quística recuperados de las bases de datos *Pubmed, Embase* y *Web of Science*. Se analizó la sensibilidad, especificidad, precisión y el área sobre la curva ROC de los programas. **Resultados:** Las búsquedas en las bases de datos dieron como resultado 1332 estudios, siendo 273 (20,5%) verdaderos duplicados. En comparación a los datos de referencia, el programa "Apoyo a la Revisión Sistemática" identificó mayor proporción de duplicados verdaderos que los demás. Los valores de sensibilidad, especificidad y precisión del sistema "Apoyo a la Revisión Sistemática" fueron superiores a 98%. **Conclusión e implicaciones para la práctica:** El sistema "Apoyo a la Revisión Sistemática" posee alta sensibilidad, especificidad y precisión para identificación de estudios duplicados obtenidos a partir de búsquedas en bases de datos en el área de salud, optimizando el trabajo de investigadores.

Palabras clave: Exactitud de los Datos; Bases de Datos Bibliográficas; Revisión Sistemática; Sensibilidad y Especificidad; Software.

Corresponding author:
Fernanda Martins Dias Escaldelai.
E-mail: fernandamartins@alumni.usp.br

Submitted on 05/04/2022.
Accepted on 09/05/2022.

DOI:https://doi.org/10.1590/2177-9465-EAN-2022-0143en

INTRODUCTION

In the health area, a growing number of systematic and scoping reviews are being conducted owing to their relevance for synthesizing scientific knowledge. Systematic reviews can help guide decisions in care provision and inform the devising of guides, recommendations and public health policies.[1]

One of the first stages of a systematic review is identifying relevant studies on different bibliographic databases.[2] Given that journals are often indexed on multiple databases, search results can contain many duplicate records.

De-duplication is the identification and removal of duplicate studies, a time-consuming task for reviewers. Computer-based tools are recommended for this task,[3] such as the reference managers EndNote®, Mendeley® and Zotero® or software tools such as Rayyan® and Start® which, besides this feature, also offer other specific functions for systematic reviews.[4]

The use of free or paid software can cut down the time required to perform de-duplication. A previous study reporting the completion of a systematic review in two weeks through computer-based tools required only 16 minutes to remove duplicates (n=1694 studies),[5] a considerably shorter timeframe than if performed by hand.

However, the quality of automated duplicate detection can be compromised by inconsistencies in references, such as differences or errors in the spelling of terms and missing data, requiring manual removal of remaining duplicates.[6]

The Systematic Review Support web-based system was developed to aid health professionals, researchers, undergraduates and graduate students in the initial stages of systematic reviews: detection and removal of duplicate references and study selection in the eligibility stage.[7] Testing the system´s ability to identify true duplicates can help demonstrate the degree of accuracy of the tool.

The objective of the present study was to assess the performance of the Systematic Review Support web-based system in the identification of duplicate records versus similar software tools typically used in the academic setting, for research studies, and by health professionals.

METHOD

This methodological study analyzed the results yielded by the Systematic Review Support (*Apoio à Revisão Sistemática*) web-based system, version 1.0[7] for identification of duplicate studies. The results were compared to those of the EndNote X9® and Rayyan® systems, adopting manual checking of duplicates as the benchmark reference, by reporting the sensitivity, specificity, accuracy and area under the ROC curve of the systems tested.

In the present investigation, a duplicate study was defined as a bibliographic record (authorship, title, journal, number, volume, number of pages) retrieved more than once on one or more electronic databases, irrespective of abbreviations or differences in spelling of terms. When references had the same title but missing data or typing errors in terms, such as volume or number of pages,[8] the studies were considered duplicates if they had the same abstract.

The Systematic Review Support system is based on a theoretical framework and on experience in conducting bibliographic review studies. An Information Technology (IT) professional independently and voluntarily developed the system using the Agile methodology. The user interfaces were created in Portuguese and the layout adapts to fit different display sizes. The system can be accessed using a device with an updated browser connected to the internet. The use of this technology cuts down the time required to remove duplicates, select eligible articles, and resolve disagreements, in addition to assuring both reliability and reproducibility, while reducing the workload of these stages for reviewers involved in health care and research.

Identification of duplicates using computer-based tools

The Systematic Review Support web-based system identifies duplicate articles by comparing titles and year of the imported records, considering alphanumeric characters, without distinguishing case sensitivity, spaces or special characters.

EndNote X9® is a reference management software tool which compares author, year, title and publication type to identify duplicates.[9] The online version was used because it is available free of charge and often used in systematic review studies, although not always cited in the resultant review articles.[10]

Rayyan® is a specific computer-based tool for conducting systematic reviews[11] and for which good results in the automated identification of true duplicates have been reported in the literature.[12] Unlike the other tools tested, duplicates are removed after final checking by the researcher.

Reference method

To perform the tests, we opted for studies about cystic fibrosis, a hereditary recessive genetic disorder predominantly affecting the lungs and digestion system, possibly leading to malnutrition.[13] The nutritional status of patients is fundamental, given its association with lung disease and survival of those affected. Search strings for this subject were defined using MeSH (Medical Subject Head) terms and input into the Pubmed, Embase and Web of Science databases (Chart 1). Literature reviews were not included.

The following procedures were adopted for manual identification of duplicate studies: (*i*) importing of records from the databases into EndNote X9® to produce reference lists in Vancouver format; (*ii*) transfer of each list into a spreadsheet; (*iii*) removing parentheses from titles prior to placing them in alphabetical order, if present; (*iv*) comparison of references and defining of single (non-duplicate) or duplicate status; (*v*) in cases of missing data, verification of abstract or full article was performed to determine the duplicate status of the study. The procedures were repeated for confirmation of results.

Evaluation of computer-based tools

After the automated identification of duplicate studies with the tools, the results were transferred to a spreadsheet and classified into true positive (true duplicate), false positive (single record incorrectly identified as a duplicate), false negative (duplicate incorrectly identified as single record) and true negative (true single record)[14] (Chart 2).

Statistical analysis

The analysis of sensitivity, specificity and accuracy provided an assessment of the performance of each tool. Sensitivity was defined as the program´s ability to correctly identify duplicate studies [a/(a+b)]; specificity as the ability to correctly detect single studies [d/(c+d)];[8,14] accuracy as the proportion of duplicate and single studies correctly identified by each tool relative to total studies retrieved from the 3 databases, according to topic [(a+d)/(a+b+c+d)].[14]

The duplicate identification rate is calculated by the proportion of duplicate studies correctly identified by the benchmark reference method relative to the total studies retrieved from the databases [a+b/(a+b+c+d)]. This data was used to calculate the accuracy of each system.[15]

Based on the calculation of the area under the curve (AUC) Receiver Operating Characteristic (ROC) and of the 95% confidence interval (95%CI), the performance of the systems overall for the total studies and for each individual cystic fibrosis topic was compared.

All analyses were performed using the statistical software packages Stata version 13 (Stata Corp LP, Texas, USA) and MedCalc®, online version.[15]

RESULTS

The database searches led to the retrieval of 1332 studies, comprising 569 (42.7%) from Pubmed, 545 (40.9%) from Embase and 218 (16.4%) from Web of Science. Of these total records, 273 (20.5%) were true duplicates (Table 1).

Compared against the benchmark reference data (Table 2), the Systematic Review Support tool had a higher duplicate detection rate than both EndNote X9® and Rayyan® systems for the three cystic fibrosis data sets: for topic 1, 98.4% (n=189) versus 59.4% (n=114) and 96.4% (n=185); for topic 2, 100% (n=38) versus 57.9% (n=22) and 97.4% (n=37); and, for topic 3, 100% (n=43) versus 65.1% (n=28) and 95.4% (n=41). For all three topics, the sensitivity of the Systematic Review Support tool proved to be higher than the other two systems tested. For topic 1 (n=834), the specificity of the system was 99.8% owing to a false positive result.

The areas of the ROC curves for the total studies (n=1332) were 0.9940 (95%CI: 0.988-1) for Systematic Review Support, 0.8004 (95%CI: 0.771-0.829) for EndNote X9® and 0.9812 (95%CI: 0.970-0.992) for Rayyan® (Figure 1). The results of the analyses, according to cystic fibrosis topic, are shown in Table 3.

DISCUSSION

The results of this study confirmed high sensitivity and specificity of the Systematic Review Support system for duplicate detection.

Chart 1. Search strings for studies on cystic fibrosis input into Pubmed, Embase and Web of Science electronic databases.

Topic	String	Period
1.Anthropometry and body composition measures and indices	cystic fibrosis AND (child OR adolescent) AND (nutrition assessment OR nutritional status OR body composition OR anthropometry OR Absorptiometry, Photon OR electric impedance OR electric conductivity OR body mass index OR waist circumference OR skinfold thickness OR body weight OR body height) *	2008 - 2018
2. Body mass index and lung disease	cystic fibrosis AND (child OR adolescent) AND body mass index AND lung diseases	2011 - 2021
3.Nutritional status and survival	cystic fibrosis AND (child OR adolescent) AND (nutritional status OR body mass index) AND survival	2011 - 2021

*__Source:__ elaborated by the authors. Search adapted from unpublished research by the authors

Chart 2. Definitions of results of systems regarding duplicate identification.

Result of system	Benchmark reference	
	Positive	Negative
Positive	(a) Duplicate flagged by system as duplicate (True Positive)	(c) Non-duplicate, but flagged by system as duplicate. (False Positive)
Negative	(b) Duplicate, but flagged by system as non-duplicate (False Negative)	(d) Non-duplicate flagged by system as non-duplicate or single (True Negative)

Source: elaborated by the authors

Table 1. Benchmark reference data, according to cystic fibrosis topic.

Benchmark reference data	Topic 1*	Topic 2**	Topic 3***	Total
	n (%)	n (%)	n (%)	n (%)
Duplicate	192 (23.0)	38 (12.9)	43 (21.1)	273 (20.5)
Single	642 (77.0)	256 (87.1)	161 (78.9)	1059 (79.5)
Total	834 (100)	294 (100)	204 (100)	1332 (100)

*Anthropometry and body composition measures and indices. **Body mass index and lung disease. ***Nutritional status and survival
Source: elaborated by the authors

Table 2. Sensitivity, specificity and accuracy of Systematic Review Support, EndNote X9® and Rayyan® systems for duplicate identification on cystic fibrosis topics.

	Topic 1*			Topic 2**			Topic 3***		
	SR Support	*EndNote X9®*	*Rayyan®*	SR Support	*EndNote X9®*	*Rayyan®*	SR Support	*EndNote X9®*	*Rayyan®*
True positives (n)	189	114	185	38	22	37	43	28	41
False negatives (n)	3	78	7	0	16	1	0	15	2
Sensitivity (%)	98.4	59.4	96.4	100	57.9	97.4	100	65.1	95.4
False positives (n)	1	0	0	0	0	1	0	0	0
True negatives (n)	641	642	642	256	256	255	161	161	161
Specificity (%)	99.8	100	100	100	100	99.6	100	100	100
Accuracy	99.6	91.7	99.3	100	91.4	99.2	100	92.9	99.1

SR: Systematic Review: *Anthropometry and body composition measures and indices; **Body Mass Index and lung disease; ***Nutritional status and survival.
Source: elaborated by the authors

Table 3. Area under ROC curve (AUC) for each system tested, according to cystic fibrosis topic.

System	Theme 1*		Theme 2**		Theme 3***	
	AUC	95%CI	AUC	95%CI	AUC	95%CI
Systematic Review Support	0.9914	0.983-1	1	****	1	****
EndNote X9®	0.7969	0.762-0.832	0.7895	0.709-0.869	0.8256	0.754-0.898
Rayyan®	0.9818	0.968-0.995	0.9849	0.959-1	0.9767	0.945-1

*Anthropometry and body composition measures and indices; **Body Mass Index and lung disease; ***Nutritional status and survival; **** Perfect performance
Source: elaborated by the authors.

The Systematic Review Support system (version 1.0) closely mirrored the deduplicating results of the Rayyan® system, a tool widely used in the academic setting.[16] Compared to EndNote X9®, the Systematic Review Support system yielded a higher proportion of true duplicates for the three topics, leaving fewer studies incorrectly identified as single (false negatives) for manual verification by the reviewer. This outperformance was previously reported in a study assessing the "Systematic Review Assistant-Deduplication Module" (SRA-DM) versus EndNote, using records from a respiratory study. Based on four data sets, the mean percentage of duplicates found by the SRA-DM was reported as 42.8% greater than the rate detected by the EndNote tool.[8]

The Systematic Review Support system had one false positive result, for topic 1. This was a letter to the Editor whose record had differences in the authorship, volume, number and pages fields. If the algorithm had encompassed additional fields, then the classification might have been correct. The Rayyan® system yielded one false positive result, for topic 2 only. Previous studies[17,18] based on reviews including over 1000 titles also reported the identification of false positives by most computer tools

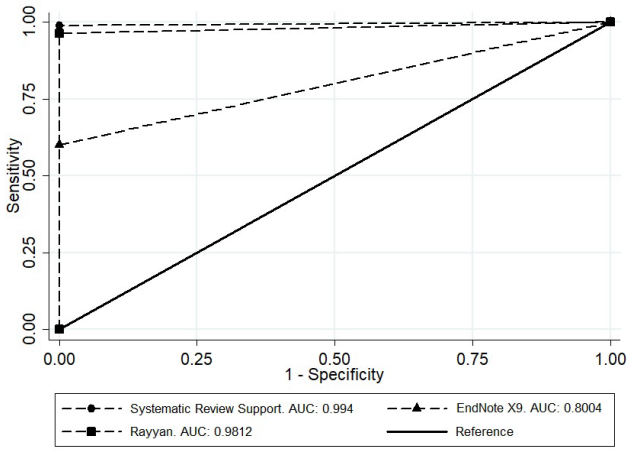

Figure 1. Area under ROC curve (AUC) for the set of studies, according to the evaluated software tool.

evaluated, including Rayyan® and EndNote X9®.[12] The occurrence of false positives is a critical issue, given that exclusion of valid studies may introduce significant selection bias in systematic reviews if these records are not later reincluded by the researcher.

The Systematic Review Support system, which incorporates an algorithm comparing title and year plus other rules, provided greater sensitivity and specificity than EndNote X9®, which uses a larger number of fields. Given the use of the year field has less variability and is deemed reliable and partially decisive,[17] its association with the title and other rules employed in the algorithm may have contributed to the high level of performance of the Systematic Review Support system, constituting a strength of the present study.

CONCLUSION AND IMPLICATIONS FOR PRACTICE

The Systematic Review Support system provided a high level of sensitivity, specificity and accuracy in the identification of duplicate studies retrieved by searches on health-related databases, optimizing the time and effort of reviewers involved in health care and research.

Representing a limitation, version 1.0 of the Systematic Review Support system processes records from only three databases; further validation tests will be thus required following an expansion to include other databases in future versions. The results of this study depend on the algorithms used by the systems and are therefore subject to change, calling for repeated comparative assessments of the tools over time.

The analyses were performed by one researcher only and the procedures were repeated for result reliability. A cystic fibrosis dataset was employed due to its interest for the researchers. Other health themes may be selected for further tests.

Although the system is dedicated to systematic reviews, the evaluated functions can be used for conducting diverse types of reviewing studies, making it useful for researchers, students, and health service professionals.

High performance computer-based tools for deduplicating can raise the quality of systemic reviews, reducing the time taken to conduct reviews. The algorithm used by the Systematic Review Support system for identifying duplicates can be enhanced in future versions.

AUTHOR'S CONTRIBUTIONS

Study design. Fernanda Martins Dias Escaldelai. Leandro Escaldelai. Denise Pimentel Bergamaschi.

Data collection or production. Fernanda Martins Dias Escaldelai. Leandro Escaldelai. Denise Pimentel Bergamaschi.

Data analysis Fernanda Martins Dias Escaldelai. Leandro Escaldelai. Denise Pimentel Bergamaschi.

Interpretation of results. Fernanda Martins Dias Escaldelai. Leandro Escaldelai. Denise Pimentel Bergamaschi.

Article writing and critical review. Fernanda Martins Dias Escaldelai. Leandro Escaldelai. Denise Pimentel Bergamaschi.

Approval of the final version of the article. Fernanda Martins Dias Escaldelai. Leandro Escaldelai. Denise Pimentel Bergamaschi.

Responsibility for all aspects of the content and the integrity of the published article. Fernanda Martins Dias Escaldelai. Leandro Escaldelai. Denise Pimentel Bergamaschi.

ASSOCIATED EDITOR

Rodrigo Nogueira da Silva

SCIENTIFIC EDITOR

Ivone Evangelista Cabral

REFERENCES

1. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol. 2018;18(1):143. http://dx.doi.org/10.1186/s12874-018-0611-x. PMid:30453902.
2. Egger M, Smith GD, Altman DG. Systematic reviews in health care: meta-analysis in context. London: BMJ Publishing Group; 2001. http://dx.doi.org/10.1002/9780470693926.
3. Lefebvre C, Glanville J, Briscoe S, Littlewood A, Marshall C, Metzendorf MI et al. 4.S1 Technical Supplement to Chapter 4: Searching for and selecting studies [Internet]. 2020 [cited 4 May 2022]. Available from: https://training.cochrane.org/handbook/current/chapter-04-technical-supplement-searching-and-selecting-studies#section-4-3
4. Kohl C, McIntosh EJ, Unger S, Haddaway NR, Kecke S, Schiemann J et al. Online tools supporting the conduct and reporting of systematic reviews and systematic maps: a case study on CADIMA and review of existing tools. Environ Evid. 2018;7(1):8. http://dx.doi.org/10.1186/s13750-018-0115-5.
5. Clark J, Glasziou P, Del Mar C, Bannach-Brown A, Stehlik P, Scott AM. A full systematic review was completed in 2 weeks using automation tools: a case study. J Clin Epidemiol. 2020;121:81-90. http://dx.doi.org/10.1016/j.jclinepi.2020.01.008. PMid:32004673.
6. 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. http://dx.doi.org/10.1371/journal.pone.0071838. PMid:23977157.

7. Revisão Sistemática. Systematic review support [Apoio à revisão sistemática] [Internet]. 2022 [cited 4 May 2022]. Available from: www.revisaosistematica.com.br
8. 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. http://dx.doi.org/10.1186/2046-4053-4-6. PMid:25588387.
9. Reuters T. EndNote X9: quick reference guide [Internet]. 2022 [cited 4 May 2022]. Available from: https://support.clarivate.com/Endnote/servlet/fileField?entityId=ka14N000000EcsXQAS&field=CA_Attachment_1__Body__s
10. Lorenzetti DL, Ghali WA. Reference management software for systematic reviews and meta-analyses: an exploration of usage and usability. BMC Med Res Methodol. 2013;13(1):141. http://dx.doi.org/10.1186/1471-2288-13-141. PMid:24237877.
11. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev. 2016;5(1):210. http://dx.doi.org/10.1186/s13643-016-0384-4. PMid:27919275.
12. 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. http://dx.doi.org/10.1186/s13643-021-01583-y. PMid:33485394.
13. Egan ME, Greene DM, Voynow JA. Fibrose cística. In: Kliegman RM, Stanton BF, St Geme II JW, Schor NF, Behrman RE, editores. Nelson: tratado de pediatria. 20ª ed. Rio de janeiro: Elsevier; 2018. p. 2098-112. (vol. 2).
14. Medronho RA, Bloch KV, Luiz RR, Werneck GL. Epidemiologia. 2ª ed. São Paulo: Atheneu; 2009.
15. MedCalc [Internet]. 2022 [cited 4 May 2022]. Available from: https://www.medcalc.org/calc/diagnostic_test.php
16. Rayyan [Internet]. 2022 [cited 4 May 2022]. Available from: https://www.rayyan.ai/
17. Jiang Y, Lin C, Meng W, Yu C, Cohen AM, Smalheiser NR. Rule-based deduplication of article records from bibliographic databases. Database. 2014 jan;2014:bat086. http://dx.doi.org/10.1093/database/bat086. PMid:24434031.
18. Kwon Y, Lemieux M, McTavish J, Wathen N. Identifying and removing duplicate records from systematic review searches. J Med Libr Assoc. 2015 out;103(4):184-8. http://dx.doi.org/10.3163/1536-5050.103.4.004. PMid:26512216.