Usefulness of machine learning softwares to screen titles of systematic reviews: a methodological study

Ana Helena Salles dos Reis[1,2], Ana Luiza Miranda de Oliveira[1], Carolina Fritsch[3], James Zouch[2], Paulo Ferreira[2†] and Janaine Cunha Polese[1*†]

Abstract

Objective To investigate the usefulness and performance metrics of three freely-available softwares (Rayyan®, Abstrackr® and Colandr®) for title screening in systematic reviews.

Study design and setting In this methodological study, the usefulness of softwares to screen titles in systematic reviews was investigated by the comparison between the number of titles identified by software-assisted screening and those by manual screening using a previously published systematic review. To test the performance metrics, sensitivity, specificity, false negative rate, proportion missed, workload and timing savings were calculated. A purposely built survey was used to evaluate the rater's experiences regarding the softwares' performances.

Results Rayyan® was the most sensitive software and raters correctly identified 78% of the true positives. All three softwares were specific and raters correctly identified 99% of the true negatives. They also had similar values for precision, proportion missed, workload and timing savings. Rayyan®, Abstrackr® and Colandr® had 21%, 39% and 34% of false negatives rates, respectively. Rayyan presented the best performance (35/40) according to the raters.

Conclusion Rayyan®, Abstrackr® and Colandr® are useful tools and provided good metric performance results for systematic title screening. Rayyan® appears to be the best ranked on the quantitative and on the raters' perspective evaluation. The most important finding of this study is that the use of software to screen titles does not remove any title that would meet the inclusion criteria for the final review, being valuable resources to facilitate the screening process.

Keywords Citation screening, Text mining, Machine learning, Software tools, User Experience

†Paulo Ferreira and Janaine Cunha Polese are share senior authorship.

*Correspondence:
Janaine Cunha Polese
janaine.polese@cienciasmedicasmg.edu.br
[1] Post-Graduate Program of Health Sciences, Faculdade Ciências Médicas de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil
[2] Faculty of Health Sciences, The University of Sydney, Sydney, NSW, Australia
[3] Faculty of Medicine and Health, School of Health Sciences, Sydney Musculoskeletal Health, The Kolling Institute, The University of Sydney, Sydney, NSW, Australia

What is new?

- Key findings

There are multiple machine learning tools that reviewers can use to facilitate and accelerate the title screening process while maintaining the quality of systematic reviews. The machine learning algorithms use reviewers' relevance labels, keywords and text mining to predict which of the titles are relevant for the study.

dos Reis *et al. Systematic Reviews*   (2023) 12:68

Page 2 of 14

- What this adds to what is known?

This study reported on the usefulness, performance metrics, and researcher's experience of three different machine learning softwares to semi-automated title screening process for systematic reviews.

- What is the implication, what should change now?

Rayyan®, Abstrackr® and Colandr® are useful, sensitive, and specific softwares to screen titles in systematic reviews, and can be safely used for title screening to save the workload and time of researchers. Overall, Rayyan® provided the best scores in the objective evaluation and on the raters' perspectives.

Background

Evidence-based practice is the pillar of decision-making for health professionals and integrates professional theoretical-practical knowledge, individual patient preferences, and high quality scientific evidence available in the literature [10, 14]. However, the search for the best literature for implementation in clinical practice is a complex process [17], especially considering the increase of over 2000% in the number of published studies over the last 20 years [11]. Furthermore, thousands of articles are published every year worldwide on different subjects and standards of care with varied levels of quality [23].

Systematic reviews are studies of high methodological quality developed through rigorous research processes that provide a reliable and valid summary of the available evidence on a specific subject, such as health interventions [6]. Systematic reviews are considered a reliable source to evaluate the quality and efficacy of health interventions, and improving the speed of development of these reviews is key to support evidence-based practice [1, 18].

Recently, several tools have been made available to speed up and facilitate the systematic review process. These tools help to reduce the costs needed to develop systematic reviews by decreasing manual work and time commitment of researchers through machine learning and text mining [1, 5]. Machine learning and text mining are features that refer to how softwares can learn through experience and repetition [5], while searching for keywords previously established by the researcher. Some softwares have been used to accelerate the process of a systematic review through semi-automatic screening of titles [3, 13]. However, the measurement properties and performance metrics of these softwares are still unknown, therefore they must be evaluated to help researchers understand if the softwares is suitable for title screening in systematic reviews when compared to

the manual screening and which of the softwares is the most appropriate for this matter [18]. Therefore, the aim of this study was to investigate the usefulness and performance metrics of three freely-available softwares (Rayyan®, Abstrackr® and Colandr®) for title screening of systematic reviews.

Methods
Design
This was a methodological study that aimed to investigate the usefulness and performance metrics of three softwares (Rayyan®, Abstrackr® and Colandr®) for the title screening process of systematic reviews.

Procedures
To be included and evaluated in the current study, softwares had to fulfill the following criteria: (1) to be freely available; (2) to address health or multidisciplinary disciplines; (3) to use text mining or machine learning tools; and (4) to assist the process of screening of titles for systematic reviews. Subsequently, we used the study published by Harrison et al. [9] to support the selection of three highest-ranked softwares with the best scores on availability and screening features (Fig. 1). The three best-evaluated softwares, which were included in the current study, were: Rayyan®, Abstrackr® and Colandr®. Appendix 1 describes the characteristics of each selected software.

The screening process was initiated through manual screening and was followed by the software-assisted screening with an interval of three to seven days between screenings to avoid memory bias. The usefulness and performance metrics of the softwares were evaluated comparing the titles and the number of titles found on the manual screening and on the software-assisted screening. For the title screening process, manual screening was considered the 'gold standard' for comparison [12].

In the screening process, three raters, R.1 (AH), R.2 (CF) and R.3 (AL), with different experiences with the use of softwares and the systematic review process were chosen and allocated to use the softwares they had not had previous experience with. Each software was used by two different raters and a fourth rater (R.4, JP), who had the most expertise in the systematic review process, resolved all conflicts. After screenings, a short survey based on the Delphi technique [2] was developed to evaluate the rater's subjective experiences regarding the softwares' performances, as described in Appendix 2. The survey consisted of four questions concerning the process of learning how to use the softwares and their user-friendly and time-saving characteristics. Each question had three to five choice answers that ranged in scores from 0 to 10, with the final grade being the average of the

dos Reis *et al. Systematic Reviews*   (2023) 12:68

Page 3 of 14
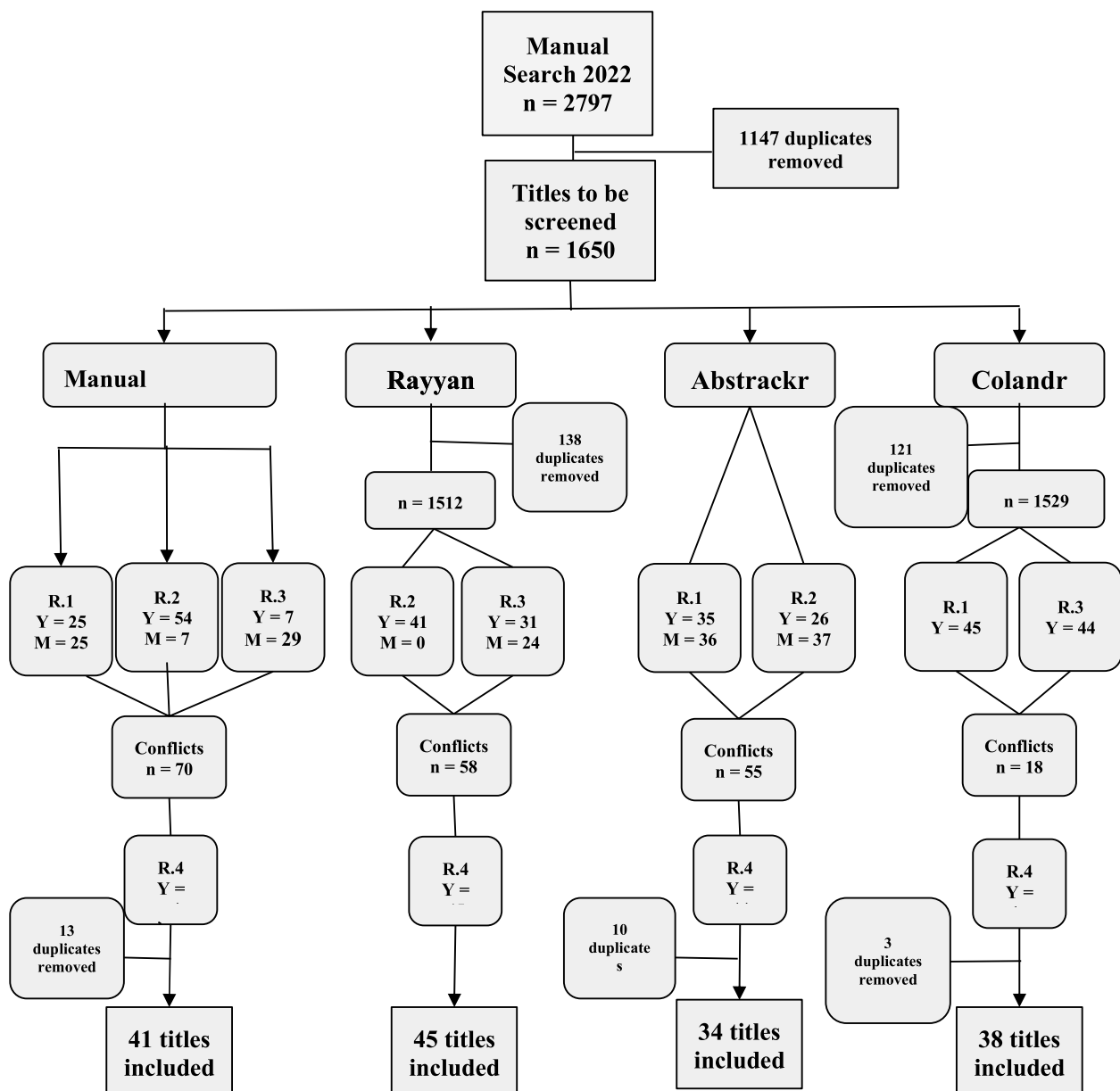
Fig. 1 Flowchart of manual and software-assisted screening process. Description: Y = yes (titles that were included after the screening); M = maybe (titles that the rater had doubts whether to include or exclude); conflicts = when raters disagreed on the inclusion or exclusion titles; R.1 (AH); R.2 (CF); R.3 (AL); R.4 (JP)

two raters' scores. This survey was pilot tested, updated and approved by 10 experts in the rehabilitation research field. The first three raters answered the questionnaire and their answers were analyzed by R.4, who summarized the scores of each software.

A systematic review in the musculoskeletal field, previously conducted and published by our research group [7], was chosen as the basis for the title screenings of the current study. The review investigated the effects of

family-based interventions compared to individual-only interventions on pain intensity and disability of people with musculoskeletal pain. A total of 18 randomized controlled trials were included in the review following a manual screening process conducted by two independent researchers. The published search strategy is presented in Attachment A and was used to conduct the search strategy on Medline, Embase, PsycINFO, Amed, Web of science, PEDro and Cinahl databases. EndNote, a citation

dos Reis *et al. Systematic Reviews*   (2023) 12:68

Page 4 of 14

management software, was used to reduce the workload removing duplicates records using EndNote deduplication system and to download all records to each software via Research Information Systems' files.

Statistical analysis

The number of articles and the titles found through the manual screening and software-assisted screening processes were compared using a percentage agreement between them. As described by Valizadeh et al. [24], a performance metric assessment was conducted considering the values of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) of each software. These results were manually calculated by the researchers. Regarding the included and excluded titles on each screening, the following formulas described by the same authors were used to calculate the metrics:

– Sensitivity = TP / (TP + FN).
– Specificity = TN / (TN + FN).
– Precision = titles correctly identified as relevant / all titles identified as relevant.
– False negative rate = titles incorrectly identified as irrelevant / all titles identified as relevant.
– Proportion missed = titles incorrectly predicted as irrelevant / all titles predicted as irrelevant.
– Workload savings = titles predicted as irrelevant / total titles to be screened.
– Time saving* = [(titles predicted as irrelevant × 0.5) / 60] / 8.

*As described by Valizadeh et al. [24], the estimated time saving is based on a screening rate of 0.5 min per title and an 8-h workday.

To describe raters' subjective experiences with Rayyan®, Abstrackr® and Colandr® softwares, a descriptive analysis was used.

Results

The current manual search process was carried out on the selected databases and identified 2797 titles, with 1650 remaining following the exclusion of the duplicates. As described on Fig. 1, all three raters firstly conducted a manual screening of titles on Excel and defined each title as included, excluded or maybe. Then, the fourth rater (R.4) resolved the conflicts and included 41 titles. Subsequently, two raters independently and blindly screened the titles on each software and each rater performed assisted screenings on two different softwares.

Raters R.2 and R.3 were responsible for the screening on Rayyan®, and each included 41 and 31 titles, respectively, and only R.3 selected maybe's (*n* = 24). As Rayyan® detected the duplicates early on the process (*n* = 138),

at the end of the screening, 58 titles were in conflict and 45 were included with no need to remove any more duplicates. R.1 and R.2 performed the screening on Abstrackr®, including 35 and 26 titles, each, with 36 and 37 maybe's respectively. 55 titles were in conflict, and after R.4's analysis and the late removal of 10 duplicates, 34 titles were included. Finally, R.1 and R.3 included 45 and 44 titles on Colandr®, with no maybe's as the software does not allow this selection. It automatically removed 121 duplicates at the beginning of the screening, but after conflict resolution (*n* = 18), it required the late removal of three duplicates, totalling 38 included titles.

In the primary review [7], researchers originally performed a manual search on the databases and found 1634 titles. After duplicates removal, 1223 titles were considered for title screening and eight extra citations were included after manual search. 18 titles were included in the review at the end of the screening process, being seven of them titles identified through manual search. Three out of the 18 titles were not found through the current databases searches and were therefore not included in the screening process. Thus, an average of 93.3% of the 15 titles included in both the primary review [7] and the current manual search were also included during the software-assisted title screenings on all three softwares (Fig. 2).

In the manual screening results, raters included 93.3% (14 titles) of the titles included in the primary review [7]. On Rayyan®, 100% were included (15 titles), while on Abstrackr® and Colandr® 93.3% (14 titles) and 86.6% (13 titles) were included, respectively, as described on Appendix 3.

Rayyan® was the most sensitive software, with the biggest proportion of titles correctly classified as relevant by raters (78%) using the software compared to those considered relevant on the gold standard method. Colandr® and Abstrackr® had sensitivity values of 65% and 60%, respectively. Rayyan® also had the lowest proportion of missed titles and false negative rate, scoring 0.5% and 21% respectively, while Colandr® scored 0.8% and 34%, and Abstrackr® scored 0.9% and 39%. All softwares had the same results for specificity, workload and time savings, indicating that semi-automated screening performed on the three softwares was faster than the manual screening (Table 1). Abstrackr® was the most precise software, followed by Rayyan® and Colandr®.

After the screenings, all three raters answered the survey about the softwares performances (Table 2). The highest ranked software was Rayyan®, with a mean score of 35 out of 40 points, followed by Abstrackr® and Colandr® which scored 33.75 and 16.25 points, respectively. Even though it took over an hour for raters to learn how
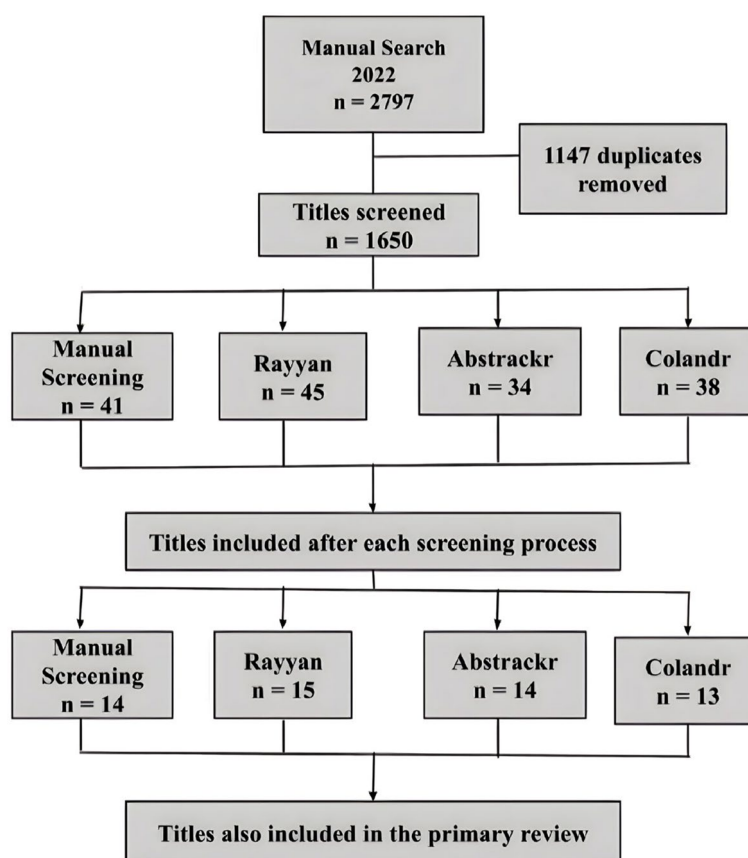
Fig. 2 Flowchart of the comparison between the included titles on the first-level title screen and the final articles included in the review

Table 1 Performance metrics of each software

Performance metric	Softwares		
	Rayyan	Abstrackr	Colandr
Sensitivity (SEN)	0.78	0.60	0.65
Specificity (SPE)	0.99	0.99	0.99
Precision	0.71	0.73	0.71
False negative rate (FNR)	0.21	0.39	0.34
Proportion missed	0.005	0.009	0.008
Workload saving	0.97	0.97	0.97
Time saving	1.67 days	1.68 days	1.67 days

to use Colandr®, which was the least intuitive one, all softwares provided faster semi-automatic screening processes when compared to manual screening.

Discussion

Results from this study revealed that Rayyan®, Abstrackr® and Colandr® are useful tools for title screening in systematic reviews, although they all had different advantages and disadvantages. A very relevant result

of this study is that the title screening softwares do not exclude any title that would meet the inclusion criteria for the final review, thus they are valuable resources to facilitate and accelerate the initial screening process. All softwares demonstrated adequate values of the investigated metrics, especially Rayyan®, which had the highest evaluation by raters. Rayyan® was the most sensitive software and raters were able to correctly identify 78% of titles deemed relevant on the manual screening, while Abstrackr® and Colandr® were sensitive to 60% and 65% of the titles, respectively. All softwares were specific, as raters correctly identified as irrelevant 99% of the titles deemed as irrelevant on the manual screening, and presented with precision of at least 71%. They incorrectly predicted as irrelevant less than 40% of the titles identified as relevant on the manual screening and were able to correctly predict as irrelevant 97% of all titles screened in all the softwares. Finally, the softwares were able to save 1.7 workdays, on average, "based on the citations that would not need to be screened" [4].

These findings mean that the softwares were able to predict and order the titles in terms of relevance, which sped up the screening processes since it was easier to

dos Reis *et al. Systematic Reviews* (2023) 12:68

Page 6 of 14

Table 2 Results of the survey on the softwares' performances according to each researcher

QUESTIONS - Softwares performance	Rayyan		Abstrackr		Colandr	
	R.2	R.3	R.1	R.2	R.1	R.3
1- How would you rate the process of learning how to use the software?	10	7.5	7.5	10	2.5	2.5
2- In your experience, would you describe the software as user friendly (i.e. is it possible to understand the software instantly and intuitively)? On a scale of 0 to 10, where 0 is not friendly at all and 10 is very friendly, how would you rate it?	10	7.5	5	10	5	0
3- How long did it take you to learn how to use the software in minutes?	7.5	7.5	5	10	2.5	0
4- Was the semi-automatic screening process faster than the manual screening process?	10	10	10	10	10	10
Total score:	35		33.75		16.25	

include and exclude titles. This acceleration of the title screening process is necessary to keep up with the exponential growth of new publications [8]. Although all three softwares had good metrics values, relying exclusively on their predictions of relevancy may result in the oversight of relevant titles [22].

The differences between the raters regarding inclusion and exclusion of titles could be possibly explained by their different levels of experience with research. R.1 participated as research assistant in four systematic reviews processes, while R.2 had previously worked in seven systematic reviews and R.3 had no previous experience with title screening. However, this difference between raters' inclusions and exclusions reduced in each screening as all of them learned the search strategies better throughout time. This was a strength of the current study as, despite their different levels of experience, they were still able to learn how to use the softwares.

Regarding the raters' evaluations, Rayyan® was the most user friendly software, ranked in the performance questionnaire with 35 out of 40 points, agreeing with the findings of Ouzzani, et al. [19]. Raters took less time to learn how to use and navigate through the software's interface, as well as to understand all of its features, as it is very intuitive. Rayyan® allows the user to choose terms for inclusion and exclusion, which might be highlighted on the titles to facilitate the screening, and it is very sensitive to detect duplicates [15, 19]. It also automatically updates the titles order by relevance every few inclusions and exclusions, as the machine learning process progresses. The software separates titles in different sections for the included, excluded, maybe's and conflicts. One limitation identified was titles that were included after conflict resolution, remained on the conflict session and had to be counted separately. Conflicts could be seen by all raters only after blinding mode was off. Finally, Rayyan® was the only software which raters identified 100% of the studies that should be included on the screening, surpassing the gold standard, manual screening.

Abtrackr® was also highly rated in the questionnaire, ranked as the second best software, and raters could

correctly identify 93.3% of the studies that should be included. Some of its positive features were that raters could be blinded, which is vital to reduce bias in the screening process. One researcher defined Abstrackr® as user friendly and praised the software for its good user interface with the best visibility of titles. It also did not demand a lot of time for the users to learn how to use it. Limitations identified included the absence of features to remove duplicate citations or highlights to favor text mining [4].

Colandr® was the least intuitive software and both raters required video guidance to learn how to navigate through its interface. In comparison with the other screenings, Colandr® was the least sensitive, although raters included 86.6% of the studies of the primary review. It was the only tool that required the definition of key terms and selection criteria for inclusion and exclusion of titles before the start of the screening process. Although this feature is positive as it allows the software to easily highlight these terms and to rank the titles by relevance, it also made rater's work slower when compared to the other softwares, as it required a reason for each exclusion [9]. Also, users were not allowed to classify titles as "maybe", which could reduce conflicts but might also be a reason for a larger number of incorrect inclusions and exclusions. Raters were not blinded, which increased the probability of bias as they were able to see all citations that were included, excluded and in conflict. Finally, Colandr® was sensitive to detect duplicate titles and removed them automatically, although a few had to be manually removed, making it a faster tool than the manual screening, similarly to the other softwares.

The findings of the current study show that these softwares can be used to facilitate the title screening process for systematic reviews, which are fundamental to the evolution of healthcare practices as they summarize the best available evidence [11]. In addition, the software proved to be a significant resource for not eliminating any title that should be included in the review, especially in title screening, which is usually the most tiring and time-consuming part of the process.

dos Reis *et al. Systematic Reviews*   (2023) 12:68

Page 7 of 14

However, researchers must take into account that as softwares are managed by humans, they are also susceptible to errors, such as the omission of relevant titles or the inclusion of irrelevant ones [8]. Moreover, as Rayyan® and Colandr® were both able to detect duplicate citations, it would be feasible to upload all titles directly on the softwares, with no need of a citation management software such as EndNote.

Limitations of the study

As only three softwares that met our inclusion criteria were evaluated, these findings should not be generalized to other softwares and abstract screening approaches. In addition, the results could be different with studies in a different field, with a higher number of citations to be screened or if the screening process was conducted by researchers with similar levels of experience with softwares and the systematic review process.

Another limitation is that the search strategy of only one systematic review was used to evaluate the softwares' performance metrics. Using any of the softwares for the title screening process for different search strategies of different reviews could also influence these results.

Regarding the researchers' subjective experience of the softwares' performances, the results cannot be generalized, since only three evaluators answered the questionnaire; and that the questionnaire contained only four questions on the topic. The differences between the TP and TN might be explained by human errors or biases, being unrelated to the softwares' performances. Furthermore, as the softwares are regularly updated, there may be some changes to these findings if future studies are conducted, and they could test the softwares using different reviews to increase credibility.

Conclusion

Rayyan®, Abstrackr® and Colandr® softwares are useful tools for title screening in systematic reviews and they all demonstrated adequate metrics values. Rayyan® appears to be the most sensitive software as it facilitated the identification of true positives by the raters and presented the least proportion of missed titles. When considering the practical use of the softwares, they had important differences regarding the subjective experience of raters. Rayyan® also had the highest score on the raters' evaluation and was considered the most user friendly and intuitive of the three softwares, which is a relevant characteristic to be considered to increase researchers' efficiency in conducting systematic reviews.

Appendix 1

Table 3 Characteristics of the softwares selected for the current study

Software	Description/site	Subject	Approach	Cost	Support	Score
Abstrackr	*Software* used of the title screening process of systematic reviews	Healthcare	*Text mining*	Completely free and free version available	Study selection Text analysis	Reorganizes the studies from 10% already screened
Colandr	Open access *Software de acesso aberto* to conduct the synthesis of evidence. It uses machine learning, natural language processing, and text mining functions to partially automate the location of relevant studies and extract the desired data from PDF articles	Multidiscipline	*Text mining Machine learning*	Completely free	Automatic search Study screening Data extraction	Every ten studies screened in the same review (including by multiple contributors), the software learns and the relevance ranking is updated
Rayyan	Collaborative web-based application to support conducting systematic reviews. It also includes a study screening mobile app	Multidiscipline	Visualization *Text mining Machine learning*	Completely free and free version available	Screening of studies Quality assessment Data extraction Collaboration Document management	*Rayyan* needs 50 screened titles with at least five included and five excluded to rank the relevance of the studies

Appendix 2
Softwares performance questionnaire.

1. How would you rate the process of learning how to use the software?

   a) Very easy (10 points)
   b) Easy (7.5 points)
   c) Not easy but also not difficult (5 points)
   d) Difficult (2.5 points)
   e) Very difficult (0 points)

2. In your experience, would you describe the software as user friendly (i.e. is it possible to understand the software instantly and intuitively)? On a scale of 0 to 10, where 0 is not friendly at all and 10 is very friendly, how would you rate it?

   a) Very Friendly (10 points)
   b) Above Average (7.5 points)
   c) Average (5 points)
   d) Below Average (2.5 points)
   e) Not friendly at all (0 points)

3. How long did it take you to learn how to use the software in minutes?

   a) < 15 min (10 points)
   b) 15 to 30 min (7.5 points)
   c) 30 to 45 min (5 points)
   d) 45 to 60 min (2.5 points)
   e) > 60 min (0 points)

4. Was the semi-automatic screening process faster than the manual screening process?

   a) Yes, with both softwares. (10 points)
   b) Yes, but with just one of the softwares. (5 points)
   c) No, both screening processes were similar regarding the time. (0 points)

Maximal score: 40 points.
Minimal score: 0 points.

Appendix 3

Table 4 Comparison of studies included in the primary review with those included in each screening process conducted in the current study

Studies included in the primary review	Included on Manual Screening	Included on Rayyan	Included on Abstrackr	Included on Colandr
Abbasi et al. 2012	Yes	Yes	Yes	Yes
Kole-Snijders et al. 1999	No	No	No	No
Saarijarvi et al. 1991	Yes	Yes	Yes	Yes
Saarijarvi et al. 1991	Yes	Yes	Yes	Yes
Saarijarvi et al. 1992	Yes	Yes	Yes	No
Turner et al. 1990	No	No	No	No
Buchanan et al. 2017	No	No	No	No
Keefe et al. 1996	Yes	Yes	Yes	Yes
Keefe et al. 1999	Yes	Yes	Yes	Yes
Keefe et al. 2004	Yes	Yes	Yes	Yes
Martire et al. 2008	Yes	Yes	Yes	Yes
Martire et al. 2003	Yes	Yes	Yes	Yes
Martire et al. 2007	Yes	Yes	Yes	Yes
Moore & Chaney 1985	Yes	Yes	Yes	Yes
Ramke et al. 2016	Yes	Yes	Yes	Yes
Radojevic et al. 1992	Yes	Yes	Yes	Yes
Riemsma et al. 2003	No	Yes	Yes	Yes
Lomholt et al. 2015	Yes	Yes	No	No

Attachment A

Systematic review search strategy: Fritsch CG, Ferreira ML, Silva AKF, Simic M, Dunn K, Campbell P, Foster NE, Ferreira PH. Family-based Interventions Benefit Individuals With Musculoskeletal Pain in the Short-term but not in the Long-Term: A Systematic Review and Meta-Analysis. Pain. 2021 Fev; Volume 37, Number 2, 140—157.

MEDLINE
1. exp Musculoskeletal Diseases/
2. exp back pain/ or exp low back pain/ or exp chronic pain/ or exp musculoskeletal pain/ or exp neck pain/
3. (low* back adj3 (pain* or ach*)).mp.
4. (low* and (backpain or backache)).mp.
5. ((lumbar or spinal vertebral) and pain*).mp.
6. (lumbago or dorsalgia).mp.
7. ((knee* or hip* or hand* or shoulder* or neck or elbow* or ankle* or wrist* or foot or feet) adj3 pain*).mp.
8. 1 or 2 or 3 or 4 or 5 or 6 or 7
9. exp couples therapy/ or exp family therapy/ or exp marital therapy/
10. ((Mother* or Father* or Brother* or Sister* or Husband* or Wife* or Household* or famil* or spous* or couple* or sibling* or offspring* or parent* or partner* or marriage* or dyad*) adj3 (interven* or trial* or therap* or counsel*)).mp.
11. 9 or 10
12. exp clinical trial/ or exp controlled clinical trial/ or exp randomized controlled trial/
13. exp clinical trials as topic/ or exp controlled clinical trials as topic/ or exp randomized controlled trials as topic/
14. random* control* trial*.mp.
15. exp Random Allocation/
16. random*.mp.
17. 12 or 13 or 14 or 15 or 16

EMBASE

1. exp musculoskeletal disease/
2. exp musculoskeletal pain/ or exp neck pain/ or exp shoulder pain/
3. exp backache/ or exp low back pain/
4. exp chronic pain/
5. (low* back adj3 (pain* or ach*)).mp.
6. (low* and (backpain or backache)).mp.
7. ((lumbar or spinal vertebral) and pain*).mp.
8. (lumbago or dorsalgia).mp.
9. ((knee* or hip* or hand* or shoulder* or neck or elbow* or ankle* or wrist* or foot or feet) adj3 pain*).mp.
10. exp osteoarthritis/ or exp arthritis/ or exp hand osteoarthritis/ or exp hip osteoarthritis/ or exp knee osteoarthritis/
11. exp fibromyalgia/
12. exp rheumatoid arthritis/
13. 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12
14. exp couple therapy/
15. exp family therapy/ (12862)
16. exp marital therapy/ (3014)

17. ((Mother* or Father* or Brother* or Sister* or Husband* or Wife* or Household* famil* or spous* or couple* or sibling* or offspring* or parent* or partner* or marriage* or dyad*) adj3 (interven* or trial* or therap* or counsel*)).mp.
18. 14 or 15 or 16 or 17
19. exp controlled clinical trial/ or exp clinical trial/ or exp controlled study/
20. exp randomized controlled trial/ or exp "randomized controlled trial (topic)"/
21. exp "controlled clinical trial (topic)"/
22. random* control* trial*.mp.
23. random*.mp.
24. 19 or 20 or 21 or 22 or 23
25. 13 and 18 and 24

PsycINFO
1. musculoskeletal diseases.mp.
2. exp back pain/ or exp chronic pain/
3. musculoskeletal pain.mp.
4. neck pain.mp.
5. (low* back adj3 (pain* or ach*)).mp.
6. (low* and (backpain or backache)).mp.
7. ((lumbar or spinal vertebral) and pain*).mp
8. (lumbago or dorsalgia).mp.
9. ((knee* or hip* or hand* or shoulder* or neck or elbow* or ankle* or wrist* or foot or feet) adj3 pain*).mp.
10. exp arthritis/ or exp rheumatoid arthritis/
11. osteoathritis.mp.
12. exp fibromyalgia/
13. 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12
14. exp couples therapy/
15. exp family therapy/
16. exp family members/
17. marital therapy.mp. or exp Marriage Counseling/
18. ((Mother* or Father* or Brother* or Sister* or Husband* or Wife* or Household* or famil* or spous* or couple* or sibling* or offspring* or parent* or partner* or marriage* or dyad*) adj3 (interven* or trial* or therap* or counsel*)).mp.
19. 14 or 15 or 16 or 17 or 18
20. exp clinical trials/
21. controlled clinical trial.mp.
22. randomized controlled trial.mp.
23. random* control* trial*.mp.
24. random*.mp. (205128)
25. 20 or 21 or 22 or 23 or 24 (210004)
26. 13 and 19 and 25 (41)

AMED
1. exp musculoskeletal disease/
2. exp backache/ or exp low back pain/ or exp back pain/ or exp musculoskeletal pain/ or exp neck pain/
3. chronic pain.mp.
4. (low* back adj3 (pain* or ach*)).mp.
5. (low* and (backpain or backache)).mp.
6. ((lumbar or spinal vertebral) and pain*).mp.
7. (lumbago or dorsalgia).mp.
8. ((knee* or hip* or hand* or shoulder* or neck or elbow* or ankle* or wrist* or foot or feet) adj3 pain*).mp.
9. exp arthritis/ or exp arthritis rheumatoid/ or exp osteoarthritis/ or exp osteoarthritis hip/ or exp osteoarthritis knee/
10. exp fibromyalgia/
11. 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10
12. couples therapy.mp.
13. exp family therapy/
14. marital therapy.mp.
15. ((Mother* or Father* or Brother* or Sister* or Husband* or Wife* or Household* or famil* or spous* or couple* or sibling* or offspring* or parent* or partner* or marriage* or dyad*) adj3 (interven* or trial* or therap* or counsel*)).mp.
16. 12 or 13 or 14 or 15
17. random* control* trial*.mp.
18. random*.mp.
19. exp clinical trials/ or exp randomized controlled trials/ or exp random allocation/
20. 17 or 18 or 19
21. 11 and 16 and 20

Web of Science
1. (pain* or ache* or backpain* or backache*)
2. ((mother* or father* or brother* or sister* or husband* or wife* or household* or famil* or spous* or couple* or sibling* or offspring* or parent* or partner* or marriage* or dyad*) near/3 (interven* or trial* or therap* or counsel*))
3. (random* control* trial*)
4. 1 and 2 and 3

PEDro
Marital therapy and musculoskeletal and clinical trial
Couple therapy and musculoskeletal and clinical trial
Family intervention and musculoskeletal and clinical trial

CENTRAL
1. exp Musculoskeletal Diseases/
2. exp musculoskeletal pain/ or exp back pain/ or exp low back pain/ or exp chronic pain/ or exp neck pain/
3. exp Shoulder Pain/
4. backache.mp.
5. (low* back adj3 (pain* or ach*)).mp.
6. (low* and (backpain or backache)).mp.
7. ((lumbar or spinal vertebral) and pain*).mp.
8. (lumbago or dorsalgia).mp.
9. ((knee* or hip* or hand* or shoulder* or neck or elbow* or ankle* or wrist* or foot or feet) adj3 pain*).mp.
10. exp arthritis/ or exp arthritis, rheumatoid/ or osteoarthritis/ or exp osteoarthritis, hip/ or exp osteoarthritis, knee/ or exp osteoarthritis, spine/
11. exp Fibromyalgia/
12. 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11
13. exp couples therapy/ or exp family therapy/ or exp marital therapy/
14. ((Mother* or Father* or Brother* or Sister* or Husband* or Wife* or Household* or famil* or spous* or couple* or sibling* or offspring* or parent* or partner* or marriage* or dyad*) adj3 (interven* or trial* or therap* or counsel*)).mp.
15. 13 or 14
16. exp clinical trial/ or exp controlled clinical trial/ or exp randomized controlled trial/
17. controlled study.mp.
18. random* control* trial*.mp.
19. exp Random Allocation/ or random*.mp.
20. 16 or 17 or 18 or 19 (1081422)
21. 12 and 15 and 20 (200)

CINHAL
1. (MH "Musculoskeletal Diseases+")
2. (MH "Knee Pain+") or (MH "Back Pain+") or (MH "Neck Pain") or (MH "Shoulder Pain")
3. (MH "Pain+") OR (MH"Back Pain+") OR (MH "Knee Pain+") OR (MH "Neck Pain") OR (MH "Chronic Pain") OR (MH "Shoulder Pain")
4. (MH "Arthritis, Rheumatoid+") or (MH "Arthritis+") or (MH "Osteoarthritis+") or (MH "Osteoarthritis, Spine+") or (MH "Osteoarthritis, Hip") or (MH "Osteoarthritis, Knee") or (MH "Osteoarthritis, Wrist") or (MH "Fibromyalgia")
5. (low* back N3 (pain* or ach*))
6. (low* and (backpain or backache))
7. ((lumbar or s inal vertebral) and pain*)
8. (lumbago or dorsalgia)
9. ((knee* or hip* or hand* or shoulder* or neck or elbow* or ankle* or wrist* or foot or feet) N3 pain*)
10. 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9
11. (MH "Couples Counseling")
12. (MH "Family Therapy")
13. ((mother* or father* or brother* or sister* or husband* or wife* or household* or famil* or spous* or couple* or sibling* or offspring* or parent* or partner* or marriage* or dyad*) N3 (interven* or trial* or therap* or counsel*))
14. 11 or 12 or 13
15. (MH "Clinical Trials+") or (MH "Randomized Controlled Trials")
16. random* control* trial*
17. random*
18. 15 or 16 or 17
19. 10 and 14 and 18

References

1. Ananidou BS, et al. Supporting Systematic Reviews Using Text Mining. Social. Science Computer Review, Manchester. 2009;27(4):509-523. https://doi.org/10.1177/0894439309332293.
2. Akins RB, Tolson H, Cole BR. Stability of response characteristics of a Delphi panel: Application of bootstrap data expansion. BMC Medical Research Methodology. 2005;5:1–12. https://doi.org/10.1186/1471-2288-5-37.
3. Bannach-Brown A, et al. Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. Syst Rev. 2019:1-12. https://doi.org/10.1186/s13643-019-0942-7.
4. Carey N, Harte M, Cullagh LM. A text-mining tool generated titleabstract screening workload savings: performance evaluation versus single-human screening. Journal of Clinical Epidemiology, Ireland. 2022:53-59. https://doi.org/10.1016/j.jclinepi.2022.05.017.
5. Das S, et al. Applications of artificial intelligence in machine learning: review and prospect. International Journal of Computer Applications. 2015. https://doi.org/10.5120/20182-2402.
6. Donato H, Donato M. Etapas na condução de uma revisão sistemática. Acta Med Port. 2019;32(3):227-235. https://doi.org/10.20344/amp.11923.
7. Fritsch CG, et al. Effects of using text message interventions for the management of musculoskeletal pain: a systematic review. Pain. 2020;161(11):2462–2475. https://doi.org/10.1097/j.pain.0000000000001958.
8. Gates A, Johnson C, Hartling L. Technology-assisted title and abstract screening for systematic reviews: a retrospective evaluation of the Abstrackr machine learning tool. Systematic Reviews. 2018. https://doi.org/10.1186/s13643-018-0707-8.
9. Harrison H, et al. Software tools to support title and abstract screening for systematic reviews in healthcare: an evaluation. BMC Med Res Methodol. 2020;20(1):7. https://doi.org/10.1186/s12874-020-0897-3.
10. Hebert R, et al. Practical Evidence-Based Physiotherapy. 2. ed. Sydney, NSW - Australia: Elsevier Health Sciences; 2011. p. 169. http://ebookcentral.proquest.com/lib/usyd/detail.action?docID=1721905. Accessed 18 June 2022.
11. Hoffmann F, et al. Nearly 80 systematic reviews were published each day: Observational study on trends in epidemiology and reporting over the years 2000-2019. J Clin Epidemiol. 2021. https://doi.org/10.1016/j.jclinepi.2021.05.022.
12. Jones-Diette J, et al. Validation of text-mining and content analysis techniques using data collected from veterinary practice management software systems in the UK. Preventive Veterinary Medicine. 2019. https://doi.org/10.1016/j.prevetmed.2019.02.015.
13. Lau J. Editorial: systematic review automation thematic series. Syst Rev. 2019;8(1):70. https://doi.org/10.1186/s13643-019-0974-z.
14. Mcevoy MP, Lewis LK, Luker J. Changes in physiotherapy students' knowledge and perceptions of EBP from first year to graduation: a mixed methods study. BMC Med Educ. 2018. https://doi.org/10.1186/s12909-018-1212-4.
15. Mckeown S, Mir ZM. Considerations for conducting systematic reviews: evaluating the performance of different methods for de-duplicating references. Systematic Reviews, Canada. 2021:1-8. https://doi.org/10.1186/s13643-021-01583-y.
16. Mokkink LB, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010;63(7):737-45. https://doi.org/10.1016/j.jclinepi.2010.02.006.
17. Moseley AM, et al. Using research to guide practice: the physiotherapy evidence database (PEDro). Braz J Phys Ther. 2019. https://doi.org/10.1016/j.bjpt.2019.11.002.
18. O'connor AM, et al. A question of trust: can we build an evidence base to gain trust in systematic review automation technologies? Systematic Reviews. 2019. https://doi.org/10.1186/s13643-019-1062-0.
19. Ouzzani M, et al. Rayyan - a web and mobile app for systematic reviews. Systematic Reviews. 2016. https://doi.org/10.1186/s13643-016-0384-4.
20. Pilatti LA, Pedroso B, Gutierrez GL. Propriedades psicométricas de instrumentos de avaliação: um debate necessário. 2010. https://doi.org/10.3895/S1982-873X2010000100005.
21. Portney LG. Foundations of Clinical Research: Applications to Evidence-Based Practice. 4. ed. [S. l.]: F.A. Davis Company, 2020. https://fadavispt.mhmedical.com/content.aspx?bookid=2885§ionid=243179473.
22. Rathbone J, Hoffmann T, Glasziou P. Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers. Systematic Reviews. 2015. https://doi.org/10.1186/s13643-015-0067-6.
23. National Science Board. National Science Foundation. Science and Engineering Indicators 2020: The State of U.S. Science and Engineering. NSB-2020-1. Alexandria, VA; 2020. https://ncses.nsf.gov/pubs/nsb20201/. Accessed 27 Oct 2021.
24. Valizadeh A, et al. Abstrackr screening using the automated tool Rayyan. Results of effectiveness in three diagnostic test accuracy systematic reviews, [S. l.]. 2022:1-15. https://doi.org/10.1186/s12874-022-01631-8.

Publisher's Note