


RESEARCH ARTICLE

Open Access



A new method for testing reproducibility in systematic reviews was developed, but needs more testing

Dawid Pieper^{1*} , Simone Heß¹ and Clovis Mariano Faggion Jr.²

Abstract

Background: To develop and test an approach to test reproducibility of SRs.

Methods: Case study. We have developed an approach to test reproducibility retrospectively while focusing on the whole conduct of an SR instead of single steps of it. We replicated the literature searches and drew a 25% random sample followed by study selection, data extraction, and risk of bias (ROB) assessments performed by two reviewers independently. These results were compared narratively with the original review.

Results: We were not able to fully reproduce the original search resulting in minor differences in the number of citations retrieved. The biggest disagreements were found in study selection. The most difficult section to be reproduced was the RoB assessment due to the lack of reporting clear criteria to support the judgement of RoB ratings, although agreement was still found to be satisfactory.

Conclusion: Our approach as well as other approaches needs to undergo testing and comparison in the future as the area of testing for reproducibility of SRs is still in its infancy.

Keywords: Systematic reviews, Reproducibility of Results, Methodological quality, Data extraction, Risk of bias, Information storage and retrieval

Introduction

Systematic reviews (SRs) are essential to inform evidence-based decision making in health care across different groups such as clinicians, patients and policy makers. Despite this huge importance and potentially resulting implications for patients-related outcomes, it has been argued that currently there is a massive production of unnecessary, misleading, and conflicted systematic reviews and meta-analyses [1]. Among others, the Lancet series *reducing waste in research* recommended research studies to undergo rigorous independent replication and

reproducibility checks [2]. In short, replication means that independent people will collect new data, while answering the same question. In contrast, reproducibility means that independent people will analyze the same data [3]. Given the definitions of replication and reproducibility from above, it becomes clear that replicability should be the ultimate goal and can be regarded to be placed over reproducibility. However, full and independent replication might not be feasible due to resource constraints. In this case, reproducibility can be seen as a way to serve as a minimum standard for judging scientific claims [4].

It was found that reproducible research practices are uncommon in SRs, and thus limiting the possibility of testing for reproducibility [5]. Others dealt with single steps of conducting SRs. For example, studies found the

*Correspondence: Dawid.pieper@uni-wh.de

¹ Institute for Research in Operative Medicine, Faculty of Health, School of Medicine, Witten/Herdecke University, Ostmerheimer Str. 200, 51109 Cologne, Germany

Full list of author information is available at the end of the article



© The Author(s) 2021. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

reproducibility of search strategies to be poor [6, 7]. Others indicated that study selection, data extraction, risk of bias assessments and meta-analyses might also lead to different results depending on the author group involved [8–11]. This implicates that these steps might also be difficult to be fully reproducible. Gaps in reproducibility in several steps of a SR potentially results in a lack of replicability.

Some first ideas have been presented how testing for reproducibility in SRs could work [12]. However, to the best of our knowledge no testing of the whole SR instead of single steps has been conducted. Therefore, we set out to develop and execute a strategy to test for reproducibility in a SR. Our strategy comprised the reproducibility of the following steps of a SR: search, selection, data extraction and risk of bias (RoB) assessment.

Methods

The methods section is divided into two parts. The first part (2.1) describes our developed idea for proportional testing for reproducibility in systematic reviews (PTRSR). This approach is tested on a single SR. This is described in the second part (2.2).

Proportional testing for reproducibility in systematic reviews (PTRSR)

One of the main ideas of the PTRSR is that it can be conducted at any time after a SR has been published (retrospective). This will allow for testing older SRs for reproducibility as well. At the same time, more than one reproduction of a SR can be conducted (e.g. by several author groups), and thus giving more power to the reproducibility test, when assuming that they come to the same result. Other approaches to test reproducibility could also include prospective elements (e.g. two independent pairs of researchers working in parallel).

The general idea of the PTRSR is that the formerly published SR is not reproduced in full, but only for a given proportion of it. This might increase feasibility given that obtaining funding and being rewarded in any

way might be difficult to achieve. According to Page et al. 2016 a therapeutic non-Cochrane SR includes a median of 14 included studies [13]. Thus, we suggest starting with a 25% proportion test, i.e. only 25% of the SR will undergo the reproducibility test. This would result in approximately 3.5 studies per SR what we have considered to be the minimal value allowing for a meaningful test. However, this is an arbitrary choice. This number needs to be adjusted when the SR does only include few studies. It should be noted that the 25% do refer to the number of hits obtained from the literature search, but not to the finally included number of studies.

In a first step, the reproducibility team (RT) will conduct all searches in bibliographic databases. After having merged all search results into one database the RT will compare their number of hits with the number of hits in the original review. If no major disagreements will be identified at this stage, the RT will draw a 25% random sample of all identified hits. After that, all forthcoming steps will be performed in the same way as reported in the original SR (see Table 1).

Results (i.e. the comparison of the original SR with the reproduced SR) can be categorized for different steps depending on their importance for reproducibility. In other words, different results can occur for some steps of the SR process without decreasing the overall confidence in the findings of the SR. For example, the number of full-texts to screen (i.e. abstracts that met eligibility criteria at title & abstract screening) is likely to vary between different research groups as some research groups might be more inclusive than other. However, no differences should be expected for the included studies, when clear eligibility criteria are reported. All relevant studies need to be included in both SRs. All studies from the original SR should be included by the RT, while at the same time the RT should not identify more eligible studies than in the original SR.

Table 1 Stepwise approach for applying the reproducibility concept to systematic reviews

Step 1 (obligatory): replicate all searches in bibliographic databases and combine them in one database
Step 2 (obligatory; percentage can be increased, e.g. in case of review including a small number of studies): draw a 25% random sample
Step 3 (obligatory): perform study selection (title&abstract and full-text) in the same way as reported, and applying the same criteria as reported in the original SR
Step 4 (obligatory): extract data for the main outcomes (e.g. primary outcomes, outcomes shown in the main text only (i. e. excluding supplementary materials))
Step 5 (obligatory): assess risk of bias/methodological quality as in the original SR
Step 6 (optional ^a): perform evidence synthesis as in the original SR. This might include meta-analyses (also including studies found to meet eligibility criteria not included in the original review) and applying systems for assessing the quality of evidence such as GRADE, for example

^a step 6 is optional as this is likely to need another approach than using the 25% sample

Case Study

We wanted to test our developed approach (2.1) on one SR (case study). Given that we have developed the approach, we might have been biased in choosing a SR. Therefore, we reached out to an information specialist and asked her to provide us a SR eligible for this case study. The information specialist was blinded against the aim of our study. The eligibility criteria were developed against the background that only well-reported SRs will qualify for undergoing testing reproducibility as outlined in 2.1 (Table 2).

With these criteria, the information specialist identified a SR entitled “Effects of omega-3 polyunsaturated fatty acid intake in patients with chronic kidney disease: Systematic review and meta-analysis of randomized controlled trials” (thenceforth labelled “original SR”) published online in 2019 that was chosen for our case study [14]. This review included sixty trials with 4,129 participants, searched several databases, performed risk of bias assessment applying the Cochrane risk of bias tool, and performed several meta-analyses including subgroup analyses.

We followed our stepwise approach (Table 1) from step 1 to step 5. Results for each step contain the comparison between the original SR and the RT. Data are presented quantitatively and discussed for each step.

Results

Searches in bibliographic databases

The original review searched MEDLINE, EMBASE and CENTRAL. The full search strategies were provided in the supplementary material. Time restrictions were also reported together with the number of hits obtained in each bibliographic database. All searches were restricted to the same end data in order to correspond to the original review. For all three databases, we were not able to get exactly the same results. For MEDLINE, we copied the complete search strategy. We used Ovid as an interface, while the interface was not explicitly reported in the original review. However, after having investigated the search string we came to the conclusion that it was very likely that Ovid was also used in the original review. While the original review retrieved 518 hits, we were only able to retrieve 462 or 499 (searching without any

time restriction) hits. Also for CENTRAL, the complete search strategy was copied and re-run. Again, the result was not fully reproducible. We retrieved fewer hits than the original review did (692 vs. 721). We have no explanation for this difference. The biggest difference between the original review and the RT was found for EMBASE (142 vs. 86). This difference might be explainable by many factors. First, we had to adapt the literature search from the Ovid to the Elsevier interface. Two text descriptors (*tw* and *sh*) used in the original research were not available when searching via the Elsevier interface. Both were replaced by the descriptors *ti*, *ab*, *kw*. It was also not clear whether the authors of the original review applied any restrictions to the publication type.

In total, after removing duplicates 922 hits were screened at title&abstract level in the original review. All but one hit were identified through the searches in bibliographic databases. One hit was identified through other sources. In contrast, our combined searches retrieved 855 hits after removing duplicates (Fig. 1).

Study selection

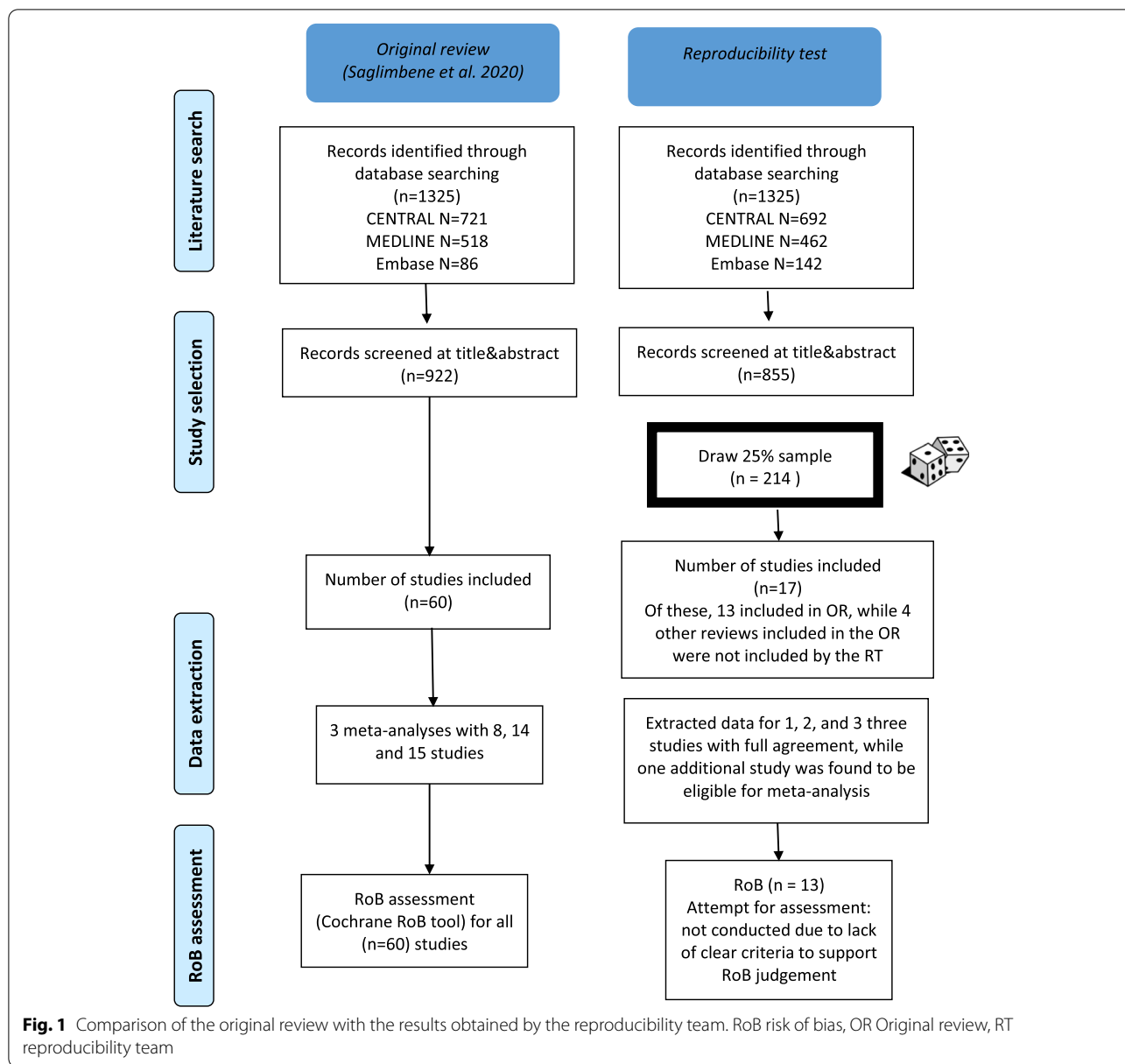
After drawing a random 25% sample, 214 citations were screened. Of those, 42 were found to be potentially relevant, and the full-text was retrieved. The RT included 17 articles meeting the eligibility criteria. Only 77% (13/17) of the full-texts included by the RT were also found to be included in the original review. No obvious reasons were found to explain the disagreement. There were also four other articles that were included in the original review, while they were found not to meet the eligibility criteria by the RT. All four articles were excluded at full-text level. Again, these disagreements could not be explained. No articles were excluded by the RT at title&abstract level that were included in the original review.

Data extraction

To check the reproducibility of data extraction, the RT checked whether the 13 studies they have found to be eligible and that were also included in the original review contained data relevant for meta-analysis. As several meta-analyses were reported in the original review, we decided to focus only on three meta-analyses. These meta-analyses focused on primary outcomes and were

Table 2 Eligibility criteria for choosing a SR for our case study

-
- SR on a healthcare intervention including only randomized controlled trials (RCTs)
 - Number of included studies ≥ 50 (to make sure that enough studies will be included and the reproducibility test will be completed successfully)
 - Search strategies reported for all bibliographic databases
 - Providing a full list of references of all included studies
 - Reported risk of bias assessment
 - Meta-analysis for at least one outcome
-



all reported and shown with a forest plot in the main article. In the case that a study was found to be eligible for meta-analysis we extracted all relevant data for the meta-analysis, i.e. number of patients and number of cases in each arm for dichotomous outcomes, and mean, standard deviation and number of patients in each arm for continuous outcomes. The meta-analyses included eight, fifteen and fourteen studies in the original review, respectively. From the 25% sample one (13%), two (13%) and three (21%) studies were included in the original review, respectively. All of these studies were identified by the RT. Furthermore, all extracted data fully matched the data in the meta-analysis in the original review.

However, we found one additional study that should have been included in the second meta-analysis (without having affected the result). This was the only disagreement in this context.

Risk of bias assessment

The RT performed RoB assessment of the 13 included RCTs. To reproduce the rating we first checked whether the authors reported the criteria used in the original SR to support the judgements of ROB. Because the SR authors did not clearly reported these criteria, we were only able to compare the ratings without taking any rationale into consideration. Two reviewers (DP, CF) of

the RT independently performed RoB assessment following the guidance in the Cochrane handbook. In a first step, we calculated raw agreement within the RT. In a second step, we compared the results of the RT against the ratings in the original SR. Four outcomes were possible at this stage: full agreement between RT and original SR (1 point), partial agreement between RT and original review (i. e. only one rater agreed with the original SR; 0.5 points), no agreement between RT and original SR (0 points). In the last case, we also aimed to differentiate whether there was agreement within the RT or not). However, there was no case of disagreement between the RT and the original SR where the RT would have also disagreed. The raw agreement from 0.65 to 0.92 and 0.38 to 0.92 depending on the outcome domain (Table 3) within the RT and between the RT and the original SR, respectively. Among the 26 disagreements among the RT only 6 disagreements were opposite (i.e. high vs. low RoB).

Discussion

To the best of our knowledge, this study is the first one to investigate the reproducibility of a SR that does not focus only on single steps, but on the whole SR. Our developed strategy to test reproducibility was based on a 25% proportion of the original review. In all investigated steps of literature search, study selection, data extraction and ROB assessment we found reproducibility to be satisfactory although some issues remained unexplained.

We were not able to fully reproduce the number of hits obtained from searching bibliographic databases. This is in line with former studies on this topic [6, 15]. To a larger extent, we found differences in study selection, including studies that should have been included according to the eligibility criteria, while other should not be included. We found only one disagreement in the extracted data that was unlikely to have an impact on the review's conclusion.

Our randomly chosen SR will probably not reflect the average reporting and methodological quality of SRs in

biomedical literature. We applied rather strict to eligibility criteria. This was necessary to secure the feasibility of our approach. The feasibility already starts with an adequate reporting of search strategies. In a sample of SRs in the field of anesthesiology only 10.5% of all SRs presented a full search strategy. Of those, only 57.4% reported the full search strategies for all sources [15]. An investigation in high impact journals found that 22% of articles provided at least one reproducible search strategy and 13% allowed reproducibility for all databases searched [6]. Older studies showed even less satisfying results [16, 17]. This might indicate that earlier calls for better reporting of search strategies have been heard [18], and reporting is improving, and thus facilitating more reproducibility checks in future. It is also important to note that we focused only on reproducing electronic searches for several reasons. Conference abstracts might be difficult and expensive to obtain if not available in an electronic format. Google scholar cannot reproduce results by definition [19]. Searching for grey literature might also be hardly reproducible. However, our choice was not only informed by pragmatism but also by evidence. Prior studies even found that searching electronic databases beyond PubMed does not lead to substantial changes in the results of SRs [20, 21]. The impact of searching grey literature may also be considered to be low, although this might be dependent on some factors such as the topic under study or the few number of studies included [21–23].

There is much less literature on study selection errors in systematic reviews. One small study estimated that pair of reviewers missed 4% of potentially relevant records, i.e. comparing pair of reviewers to the original review at title&abstract level [24]. Ford et al. replicated eight SRs of pharmacological interventions for irritable bowel syndrome and found that six meta-analyses missed 17 separate RCTs, constituting 3–11% of eligible patients not been included [25]. However, the authors did not investigate whether the reason for missing these RCTs

Table 3 Agreement in risk of bias assessment

	Random sequence generation	Allocation concealment	Blinding participants or personnel	Blinding outcome assessors	Incomplete outcome data	Selective reporting	Other Bias
Reproducibility team (within)	0.85 (11/13)	0.92 (12/13)	0.75 (15/20)	0.65 (13/20)	0.75 (15/20)	0.69 (9/13)	0.85 (11/13)
Reproduction team vs. original SR	0.92 (12/13)	0.88 (11.5/13)	0.48 (9.5/20)	0.78 (15.5/20)	0.38 (7.5/20)	0.77 (10/13)	0.62 (8/13)

could be attributed either to study selection or literature search.

Our result for data extraction was very satisfactory. Other studies also indicated that data extraction errors occur regularly leading to biased estimates [26–28], while this will usually not impact the conclusions of the SR [29, 30]. One exception was the extraction on RoB ratings from the original publication. There was limited and unclear information on the criteria used by authors to rate RoB. The full report of the rationale used to judge RoB should be a standard in any SR to allow any kind of auditing purposes. According to MECIR 2021, the rationale for RoB ratings should be reported in Cochrane intervention reviews [31]. However, this has already been discussed earlier, and should become standard in all SRs [32]. Given that usually supplementary materials can be made available online this should not be a challenge. Nevertheless, although there were disagreements between the RT and the original SR, the agreement can still be regarded to be quite satisfactory when compared to studies investigating RoB agreements between SRs [8, 9].

Overall, the first three steps (Table 1) of our approach can be well anticipated in advance. In particular step 4 (data extraction), but also steps 5 and 6, cannot be anticipated in the same way as they heavily depend on the number of included studies, outcomes and performed analyses. Thus, it is much more difficult to provide clear guidance for these steps. For example, if only one meta-analysis was performed it is a clear-cut choice that data extraction should focus on it, while the choice might be less obvious in case of multiple meta-analyses. Authors should clearly report and explain their choice for these steps.

The completeness of reporting the SR steps seems pivotal to achieve full reproducibility. For example, this is the case in the step of study selection results from the literature search. Replicating this step will hardly be ever possible. The authors would need not only to report the list of excluded studies at full-text level what is in accordance with the *Preferred reporting items for systematic reviews and meta-analyses* (PRISMA) guideline [33], but start to present a full list of all citations retrieved at title&abstract level. The steps of data extraction and RoB assessment could have been conducted independently of the others as all included studies were reported. The outcome data to retrieve were quite straightforward in our case study. We acknowledge that this is dependent on the type of data (e.g. binary vs. continuous, number of measurements (different time points)), and larger discrepancies can also be expected [34–36].

We understand that the report of every step in a SR should be complete enough to avoid any further contact

with SR authors to clarify potential issues. Further contacts would imply in lack of efficiency in the use of resources, and we feel it is not the responsibility of assessors to check data beyond of what is reported in the scientific article. In fact, we understand that it is the obligation of SR authors to provide full report to allow the reproducibility of the steps. One important question here is: how to be certain that answers provided by authors are in fact accurate? Some evidence suggests that contacting authors can modify important outcomes of SRs [37]. Thus, should we trust in the results of SRs where authors were not contacted to provide extra information?

In psychology, a new article type called Registered Replication Reports has been introduced by the journal *Advances in Methods and Practices in Psychological Science*. The RTs can submit a detailed protocol for an ongoing study. This protocol is then forwarded to the authors of the original study for feedback. In contrast to our approach, RTs already submit their plan prior to data collection of the original study. This comes along with two potential advantages. First, the result of the RT will be published irrespective of the result, and thus giving credit for their effort. Second, the described process aims more for constructive feedback than for identifying errors in others work [38]. Such an approach could potentially also be feasible for SRs in biomedicine. However, the relatively short time period between registering a review or publishing its protocol and start of data collection would give RTs only a very limited time window to submit their reproducibility proposal.

Overall, it becomes obvious that both, reporting quality and methodological quality of a SR play a crucial role, and this needs also to be kept in mind when choosing SRs for undergoing a reproducibility test. We argue that applying our approach to low quality SRs might be useless. Furthermore, critical flaws in the methods (e.g. wrong statistical methods) can be a reason not to test for reproducibility at all. Tools for assessing the methodological quality of SRs such as AMSTAR 2 [39], for example, can be applied to identify critical flaws. In some cases, however, a meta-analysis should be reproduced or conducted again in an adequate way to obtain more accurate results or clarify conflicting meta-analytic results from different meta-analyses on the same research question.

Limitations

Our randomly chosen SR focused on a health care intervention and included only RCTs. Although this is the most frequent type of a SR [13], we admit that our approach would possibly need modifications for other types of reviews (e.g. diagnostic test accuracy reviews). In general, more methods studies focusing on the single steps of conducting SRs are highly appreciated to

substantiate evidence for and reproducibility testing. A recent review revealed a lack of such studies [40]. We felt that the steps of literature search, study selection and RoB assessment were sufficient to gain knowledge of the underlying steps. When searching literature, we were not always able to use the same interfaces either due to non-reporting in the original SR or because of having no access to the interfaces used. However, we only extracted data for meta-analyses. Given the aforementioned evidence on errors in meta-analyses it might also have been interesting to reproduce one meta-analysis in full. Our approach lacks of formalized overall assessment of the PTRSR. We believe that more practical tests are needed to test the feasibility and applicability of our approach first to confirm that the domains are chosen and operationalized correctly. Last, we did not get in contact with the authors of the original review. We also have to acknowledge that the RT had no content expertise in the review under study. However, as we did not interpret the findings to come up with clinical recommendations, we think that this might be a negligible factor. Furthermore, it has been formerly argued that content expertise might not be that important for authors of SRs [41].

Conclusion

Our approach resembles a post-publication review that is performed in a structured way. Thus, reproducibility tests can become a part of such post-publication reviews and allow the original review authors to improve on their review in terms of reporting and methodological quality [42].

An essential step in reproducing SRs is that SR authors make all of their data accessible. This will allow reproducibility and increase the credibility of SRs [43]. Our approach as well as other approaches needs to undergo testing and comparison in the future as the area of testing for reproducibility of SRs is still in its infancy.

Abbreviations

RCT: randomized controlled trial; SR: systematic review; PTRSR: proportional testing for reproducibility in systematic reviews; RT: reproducibility team; RoB: risk of bias; OR: original review; PRISMA: Preferred reporting items for systematic reviews and meta-analyses.

Acknowledgements

None.

Authors' contributions

DP: Conceptualization; Methodology; Formal analysis; Investigation; Writing—Original Draft; Writing—Review & Editing; Visualization; Supervision; Project Administration. SH: Formal analysis; Investigation; Visualization; Writing—Review & Editing. CMF: Conceptualization; Methodology; Formal analysis; Investigation; Writing—Review & Editing; Visualization; Supervision. The authors read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. This study did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Availability of data and materials

All research data are available upon request from the corresponding author.

Declarations

Ethics approval and consent to participate

This study did not involve any humans.

Consent for publication

All authors provided consent for publication.

Competing interests

The authors declare to have no competing interests.

Author details

¹Institute for Research in Operative Medicine, Faculty of Health, School of Medicine, Witten/Herdecke University, Ostmerheimer Str. 200, 51109 Cologne, Germany. ²Department of Periodontology and Operative Dentistry, Faculty of Dentistry, University Hospital Münster, Münster, Germany.

Received: 21 December 2020 Accepted: 28 June 2021

Published online: 29 July 2021

References

- Ioannidis JP. The Mass Production of Redundant, Misleading, and Conflicted Systematic Reviews and Meta-analyses. *Milbank Q*. 2016;94(3):485–514.
- Moher D, Glasziou P, Chalmers I, Nasser M, Bossuyt PMM, Korevaar DA, et al. Increasing value and reducing waste in biomedical research: who's listening? *Lancet*. 2016;387(10027):1573–86.
- Patil P, Peng RD, Leek JT. A statistical definition for reproducibility and replicability. *bioRxiv*. 2016:066803.
- Peng RD. Reproducible Research in Computational Science. *Science*. 2011;334(6060):1226–7.
- Page MJ, Altman DG, Shamseer L, McKenzie JE, Ahmadzai N, Wolfe D, et al. Reproducible research practices are underused in systematic reviews of biomedical interventions. *J Clin Epidemiol*. 2018;94:8–18.
- Koffel JB, Rethlefsen ML. Reproducibility of Search Strategies Is Poor in Systematic Reviews Published in High-Impact Pediatrics, Cardiology and Surgery Journals: A Cross-Sectional Study. *PLoS One*. 2016;11(9):e0163309.
- Faggion CM Jr, Huivin R, Aranda L, Pandis N, Alarcon M. The search and selection for primary studies in systematic reviews published in dental journals indexed in MEDLINE was not fully reproducible. *J Clin Epidemiol*. 2018;98:53–61.
- Bertizzolo L, Bossuyt P, Atal I, Ravaud P, Dechartres A. Disagreements in risk of bias assessment for randomised controlled trials included in more than one Cochrane systematic reviews: a research on research study using cross-sectional design. *BMJ Open*. 2019;9(4):e028382.
- Könsgen N, Barcot O, Heß S, Puljak L, Goossen K, Rombey T, et al. Inter-review agreement of risk-of-bias judgments varied in Cochrane reviews. *J Clin Epidemiol*. 2020;120:25–32.
- Mathes T, Klaußen P, Pieper D. Frequency of data extraction errors and methods to increase data extraction quality: a methodological review. *BMC Med Res Methodol*. 2017;17(1):152.
- Waffenschmidt S, Knelangen M, Sieben W, Bühn S, Pieper D. Single screening versus conventional double screening for study selection in systematic reviews: a methodological systematic review. *BMC Med Res Methodol*. 2019;19(1):132.
- Faggion CM Jr. Should a systematic review be tested for reproducibility before its publication? *J Clin Epidemiol*. 2019;110:96.
- Page MJ, Shamseer L, Altman DG, Tetzlaff J, Sampson M, Tricco AC, et al. Epidemiology and Reporting Characteristics of Systematic

- Reviews of Biomedical Research: A Cross-Sectional Study. *PLoS Med.* 2016;13(5):e1002028.
14. Saglimbene VM, Wong G, van Zwieten A, Palmer SC, Ruospo M, Natale P, et al. Effects of omega-3 polyunsaturated fatty acid intake in patients with chronic kidney disease: Systematic review and meta-analysis of randomized controlled trials. *Clin Nutr.* 2020;39(2):358–68.
 15. Biocic M, Fidahic M, Puljak L. Reproducibility of search strategies of non-Cochrane systematic reviews published in anaesthesiology journals is suboptimal: primary methodological study. *Br J Anaesth.* 2019;122(6):e79–81.
 16. Maggio LA, Tannery NH, Kanter SL. Reproducibility of Literature Search Reporting in Medical Education Reviews. *Acad Med.* 2011;86(8):1049–54.
 17. Yoshii A, Plaut DA, McGraw KA, Anderson MJ, Wellik KE. Analysis of the reporting of search strategies in Cochrane systematic reviews. *J Med Libr Assoc.* 2009;97(1):21–9.
 18. Sampson M, McGowan J, Tetzlaff J, Cogo E, Moher D. No consensus exists on search reporting methods for systematic reviews. *J Clin Epidemiol.* 2008;61(8):748–54.
 19. Jacsó P. Google scholar revisited. Online information review. 2008.
 20. Halladay CW, Trikalinos TA, Schmid IT, Schmid CH, Dahabreh IJ. Using data sources beyond PubMed has a modest impact on the results of systematic reviews of therapeutic interventions. *J Clin Epidemiol.* 2015;68(9):1076–84.
 21. Hartling L, Featherstone R, Nuspl M, Shave K, Dryden DM, Vandermeer B. The contribution of databases to the results of systematic reviews: a cross-sectional study. *BMC Med Res Methodol.* 2016;16(1):127.
 22. Adams J, Hillier-Brown FC, Moore HJ, Lake AA, Araujo-Soares V, White M, et al. Searching and synthesising “grey literature” and “grey information” in public health: critical reflections on three case studies. *Syst Rev.* 2016;5(1):164.
 23. Hartling L, Featherstone R, Nuspl M, Shave K, Dryden DM, Vandermeer B. Grey literature in systematic reviews: a cross-sectional study of the contribution of non-English reports, unpublished studies and dissertations to the results of meta-analyses in child-relevant reviews. *BMC Med Res Methodol.* 2017;17(1):64.
 24. Edwards P, Clarke M, DiGiuseppi C, Pratap S, Roberts I, Wentz R. Identification of randomized controlled trials in systematic reviews: accuracy and reliability of screening records. *Stat Med.* 2002;21(11):1635–40.
 25. Ford AC, Guyatt GH, Talley NJ, Moayyedi P. Errors in the Conduct of Systematic Reviews of Pharmacological Interventions for Irritable Bowel Syndrome. *Official J Am Coll Gastroenterol.* 2010;105(2):280–8.
 26. Bray A, Kampouraki E, Winter A, Jesuthasan A, Messer B, Graziadio S. High Variability in Sepsis Guidelines in UK: Why Does It Matter? *Int J Environ Res Public Health.* 2020;17(6):2026.
 27. Carroll C, Scope A, Kaltenthaler E. A case study of binary outcome data extraction across three systematic reviews of hip arthroplasty: errors and differences of selection. *BMC Res Notes.* 2013;6:539.
 28. Tendal B, Higgins JPT, Jüni P, Hróbjartsson A, Trelle S, Nüesch E, et al. Disagreements in meta-analyses using outcomes measured on continuous or rating scales: observer agreement study. *BMJ.* 2009;339:b3128.
 29. Gøtzsche PC, Hróbjartsson A, Maric K, Tendal B. Data extraction errors in meta-analyses that use standardized mean differences. *JAMA.* 2007;298(4):430–7.
 30. Jones AP, Remington T, Williamson PR, Ashby D, Smyth RL. High prevalence but low impact of data extraction and reporting errors were found in Cochrane systematic reviews. *J Clin Epidemiol.* 2005;58(7):741–2.
 31. Cochrane Editorial Unit. Methodological Expectations of Cochrane Intervention Reviews (MECIR). 2021. Available from: <https://community.cochrane.org/mecir-manual>.
 32. Faggion CM Jr. The rationale for rating risk of bias should be fully reported. *J Clin Epidemiol.* 2016;76:238.
 33. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *Bmj.* 2021;372:n160.
 34. Maassen E, van Assen MALM, Nuijten MB, Olsson-Collentine A, Wicherts JM. Reproducibility of individual effect sizes in meta-analyses in psychology. *PLoS One.* 2020;15(5):e0233107-e.
 35. Mayo-Wilson E, Fusco N, Li T, Hong H, Canner JK, Dickersin K. Multiple outcomes and analyses in clinical trials create challenges for interpretation and research synthesis. *J Clin Epidemiol.* 2017;86:39–50.
 36. Mayo-Wilson E, Li T, Fusco N, Bertizzolo L, Canner JK, Cowley T, et al. Cherry-picking by trialists and meta-analysts can drive conclusions about intervention efficacy. *J Clin Epidemiol.* 2017;91:95–110.
 37. MeursingeReynders R, Ladu L, Di Girolamo N. Contacting of authors modified crucial outcomes of systematic reviews but was poorly reported, not systematic, and produced conflicting results. *J Clin Epidemiol.* 2019;115:64–76.
 38. Association for Psychological Science. Registered Replication Reports N.N. Available from: <https://www.psychologicalscience.org/publications/replication>.
 39. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *Bmj.* 2017;358:j4008.
 40. Robson RC, Pham B, Hwee J, Thomas SM, Rios P, Page MJ, et al. Few studies exist examining methods for selecting studies, abstracting data, and appraising quality in a systematic review. *J Clin Epidemiol.* 2019;106:121–35.
 41. Gøtzsche PC, Ioannidis JPA. Content area experts as authors: helpful or harmful for systematic reviews and meta-analyses? *BMJ.* 2012;345:e7031.
 42. Shokraneh F. Reproducibility and replicability of systematic reviews. *World J Meta Anal.* 2019;7(3):66–71.
 43. Lakens D, Hilgard J, Staaks J. On the reproducibility of meta-analyses: six practical recommendations. *BMC Psychol.* 2016;4(1):24.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

