Manuscript Details

Manuscript number	JCE_2020_721_R1
Title	Has the Quality of Evidence for Medical Interventions Improved? A Meta-Epidemiological Study of Cochrane Reviews
Article type	Review Article

Abstract

Background: A previous analysis of Cochrane Reviews published between January 1st, 2013 and June 30th, 2014 found that only 13.5% reported high quality evidence for the intervention according the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) system. 31.7% had low level, and 24% revealed very low level of evidence. Many of these reviews have been updated, and it is unknown whether the updated reviews report a change in the quality of evidence. Objectives: To determine the change in quality of evidence in updates of Cochrane reviews that were initially published between 1st January 2013 and 30th June 2014. Methods: We searched the Cochrane Database of Systematic Reviews on March 20th, 2020 to identify which of the reviews from the initial (2013/14) sample have been updated. Using the same methods to determine the quality of evidence in the previous analysis, we assessed the quality of evidence for the first listed primary outcomes in the updated reviews. Results: Of the 608 reviews in the original sample, 154 had been updated with 151 presenting available data for both original and updated SRs (24.8%). The updated reviews included: 15 (9.9%) with high quality evidence, 56 (37.1%) with moderate, 47 (31.1%) with low, and 33 (21.9%) with very low-quality evidence. No change in the GRADE quality of evidence was found for most (103, 68.2%) of the updated reviews. Of the 48 reviews with a change in GRADE rating (58.3%) were downgraded, mostly to low or very low. The quality of evidence rating improved in 20 (41.7%), although only 6 reviews were promoted to high quality. Conclusions: Updated systematic reviews continued to suggest that only a minority of outcomes for healthcare interventions are supported by high-quality evidence. The quality of the evidence did not consistently improve or worsen in updated reviews.

Keywords	systematic review; quality of evidence; GRADE; evidence; evidence-based medicine
Manuscript region of origin	Europe

Corresponding Author	Jeremy Howick
Corresponding Author's Institution	University of Oxford
Order of Authors	Jeremy Howick, Despina Koletsi, Nikolaos Pandis, Padhraig Fleming, Martin Loef, Harald Walach, Stefan Schmidt, John Ioannidis
Suggested reviewers	Holger Cramer, Jürgen Barth, Karin Meissner

Submission Files Included in this PDF

File Name [File Type]

In Effective Cover July 2020.doc [Cover Letter]

Reply to Reviewer Comments Final.docx [Response to Reviewers]

Cochrane GRADE Update final tracked.doc [Revised Manuscript with Changes Marked]

Highlights July 2020.docx [Highlights]

Cochrane GRADE Update final clean.doc [Manuscript File]

Declaration of interest.docx [Conflict of Interest]

Author statement.docx [Author Statement]

To view all the submission files, including those not included in the PDF, click on the manuscript title on your EVISE Homepage, then click 'Download zip file'.

Has the Quality of Evidence for Medical Interventions Improved? A Meta-Epidemiological Study of Cochrane Reviews

Jeremy Howick, PhD[1], Despina Koletsi, DiplDS, Dr. med. dent[2*], Nikolaos Pandis[3], Padhraig S. Fleming, PhD[4], Martin Loef, PhD[5], Harald Walach, PhD[5,6], Stefan Schmidt, PhD[7], John P.A. Ioannidis, MD, DSc[8]

[1] Faculty of Philosophy, University of Oxford, Oxford OX2 6GG, United Kingdom

[2] Clinic of Orthodontics and Pediatric Dentistry, Center of Dental Medicine, University of Zurich, Switzerland *joint first author

[3] Department of Orthodontics and Dentofacial Orthopedics, School of Dental Medicine, Medical Faculty, University of Bern, Bern, Switzerland

[4] Institute of Dentistry, Queen Mary, University of London

[5] CHS-Institute, Berlin, Germany

[6] Poznan University of the Medical Sciences, Department of Pediatric Gastroenterlogy, Poznan, Poland

[7] Department of Psychosomatic Medicine and Psychotherapy, Medical Center, University of Freiburg

[8] Departments of Medicine, of Epidemiology and Population Health, of Biomedical Data Science, and of Statistics, and Meta-Research Innovation Center at Stanford (METRICS), Stanford University, CA, USA

Correspondence to: Jeremy Howick, Faculty of Philosophy, University of Oxford, Oxford OX2 6GG, +44 (0)7771925412, E-mail: jeremy.howick@philosophy.ox.ac.uk

Registration

Open Science Framework: Howick, J., Koletsi, D., Fleming, P., Schmidt, S., Loef, M., Walach, H., … Ioannidis, Jj. (2020, March 30). Has the Quality of Evidence for Medical Interventions Improved? Protocol for a Meta-Epidemiological Study. Retrieved from osf.io/bw7ky

Contributions

JH, (guarantor) and JPAI conceived of the idea., JH wrote the first draft of the protocol., DK did the data extraction.; JH, ML, PF, HW checked the extraction. DK and NP did the initial analysis. All authors interpreted the analyses, contributed to drafting the protocol and writing the manuscript.

Support

The writing of this protocol was not independently funded.

Declaration of interest

None of the authors have any conflicts of interests related to this paper.

Abstract

Background: A previous analysis of Cochrane Reviews published between January 1st, 2013 and June 30th, 2014 found that only 13.5% reported high quality evidence for the intervention according the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) system. 31.7% had low level, and 24% revealed very low level of evidence. Many of these reviews have been updated, and it is unknown whether the updated reviews report a change in the quality of evidence.

Objectives: To determine the change in quality of evidence in updates of Cochrane reviews that were initially published between 1st January 2013 and 30th June 2014.

Methods: We searched the Cochrane Database of Systematic Reviews on March 20th, 2020 to identify which of the reviews from the initial (2013/14) sample have been updated. Using the same methods to determine the quality of evidence in the previous analysis, we assessed the quality of evidence for the first listed primary outcomes in the updated reviews.

Results: Of the 608 reviews in the original sample, 154 had been updated with 151 presenting available data for both original and updated SRs (24.8%). The updated reviews included: 15 (9.9%) with high quality evidence, 56 (37.1%) with moderate, 47 (31.1%) with low, and 33 (21.9%) with very low-quality evidence. No change in the GRADE quality of evidence was found for most (103, 68.2%) of the updated reviews. Of the 48 reviews with a change in GRADE rating (58.3%) were downgraded, mostly to low or very low. The quality of evidence rating improved in 20 (41.7%), although only 6 reviews were promoted to high quality.

Conclusions: Updated systematic reviews continued to suggest that only a minority of outcomes for healthcare interventions are supported by high-quality evidence. The quality of the evidence did not consistently improve or worsen in updated reviews.

Keywords: Systematic review; evidence; Quality score; Meta-analysis

What is new?

Key findings

- ~~Only a minority of Cochrane Reviews are updated within 6 years.~~
- The quality of evidence (according to GRADE) supporting the main finding changes in about a ~~third~~ quarter of updated Cochrane reviews.
- Upgrading of quality of evidence (according to GRADE) for the main outcome is not more common than downgrading of quality of evidence.

What this adds to what was known?

- Quality of evidence does not seem to improve overall with the addition of new evidence, at least within the timeframe assessed.

What is the implication and what should change now?

- Methods investigating when review updates are likely to change our confidence in the estimated outcome effect could inform decisions about whether to update reviews in order to save resources.
- The quality of evidence supporting most healthcare interventions remains low; higher quality evidence is required.

1. Introduction

1.1. Rationale

Several meta-epidemiological studies have attempted to determine the proportion of healthcare interventions that are evidence-based. A 2001 estimate found that about a quarter (26.7%) of healthcare interventions whose effectiveness was reported in 160 Cochrane Reviews were considered effective, based on the interpretation of the review authors. [1] In 2007, Garrow claimed that 50% of healthcare treatments have good evidence to support them. [2] In the same year, El Dib *et al.* (2007) found that just 44% of a random selection of Cochrane Reviews evaluating interventions suggested that they were likely to be beneficial. [3]

Since these studies were published, the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) system has been introduced offering a less subjective way of ranking the quality of evidence. [4] An evaluation of all Cochrane Reviews published between January 1, 2013 and June 30, 2014 found that 13.5% of reviews were found to have high quality of evidence for the first listed primary outcome according to GRADE.[5] High quality evidence was more common in updated compared to new reviews and in association with pharmacologic than other types of interventions. Even when any outcomes (including but not limited to the first listed primary outcome) were considered, only 116/608 (19.1%) of the reviews reported at least one outcome with high quality of evidence.

Most researchers agree that it is important to update systematic reviews so that they reflect current knowledge, [6 7] to maximize patient benefits, and to avoid harm. [8] However, updated reviews frequently reveal no change in conclusions when compared with the original. According to French et al., only about 9% of updated Cochrane Reviews in 2002 presented a change in conclusion relative to their precursors from 1998. [9] However, the claim

that the updates did not overturn results from the original review was based on whether review authors stated there was a change in the conclusion of the updated review.

There is currently no consensus on the timing that would appropriately guide a review update and the Cochrane Collaboration's policy is to update reviews when evidence accumulates, based on the availability of new data that would have a meaningful impact on the findings and on the importance of the review question. [10] Previous reports have identified a median time required for an update of a systematic review of approximately 5.5 years. [11] It was therefore considered appropriate to assess whether reviews conducted back in 2013-2014 (Fleming et al., 2016) have been updated by early 2020, and if so, whether there are changes in the quality of the evidence based on GRADE. [5]

1.2. Objectives

The primary objective was to determine whether updates from a previous sample of systematic reviews resulted in a different quality evidence, as assessed by GRADE. The secondary objectives were to determine whether there is a difference in the change of quality of evidence across different interventions, outcomes, or Cochrane Review Groups.

2. Methods

2.1. Eligibility criteria

We included any Cochrane Review that was an update of a Cochrane Review published in the (01/01/2013—30/06/2014) parent sample of reviews which included a GRADE assessment.

2.2. Information sources

Cochrane Database of Systematic Reviews: https://www.cochranelibrary.com/cdsr/reviews.

2.3. Search strategy

We searched the Cochrane Database of Systematic Reviews to identify the reviews ~~from the original sample~~ which had updates among those in the original sample. The most recent search was ~~, with last update~~ on March 20th, 2020.

2.4. Data sources and searches

One author (DK) retrieved the systematic reviews from the original (2013/14) sample and piloted the extraction form with one other author (JH). One author (DK) checked whether an update had been published and extracted data for the updated review. Other authors (JH, ML, PF, HW) were second extractors (all records were checked ~~the data extraction~~by two authors). All discrepancies were resolved by discussion.

2.5. Data items

Extracted information included: titles, corresponding author name and email, Cochrane Review Group, year of publication, country, study design, intervention (and intervention category), control and outcome. In relation to the GRADE Summary of Findings tables (SoF), ~~the following~~ we~~re~~ recorded for the first listed outcome the~~:~~ category of intervention (including surgical, pharmacologic, behavioural or medical treatments, and diet or exercise interventions). ~~In brief~~Interventions classified as:~~,~~ "behavioural" ~~interventions~~ pertained to psychological treatment, psychotherapy, cognitive training, group therapy; "diet or exercise" interventions largely related to training exercise, physiotherapy, rehabilitation, dietary modification; "medical treatments" were summarized by electronic optical/ hearing aids, appliance/ device use for dental treatment, ultrasound or other radiography and medical interventions not related to surgical or pharmacologic approaches. We also recorded~~;~~ type of outcomes (objective, such as mortality or outcomes assessed with an instrument or pre-specified measurable criteria; or subjective) and overall GRADE ranking with reasons for downgrade or upgrade. In cases where multiple Summary of Findings tables within the same review existed for the primary outcome, we considered only the one listed first. In cases where no high-quality evidence was recorded for the first listed primary outcome, we documented whether any other outcome was rated as high and, if so, whether this was a primary (but not first listed) one.

We reported whether the Cochrane review authors concluded that the experimental intervention should be used in clinical or public health practice or not. This information was obtained from the conclusions section in the review abstract and the body of the review (subsections "implications for practice" and/ or "implications for research"), following the original strategy implemented in the parent study.[5] Examples of positive interpretations were: "Buprenorphine should be supported as a medication to use," and in the "Implications for

research or practice" section: "There does not appear to be any need for further randomized control trials of the relative efficacy of methadone compared with buprenorphine."

2.6. Outcomes

The primary outcome was the change in quality of the evidence for the primary outcome in updated Cochrane Reviews compared with reviews published in an earlier (01/01/2013—30/06/2014) parent sample. The secondary outcomes were the proportion of reviews in the updated sample that have high, moderate, low, or very low-quality evidence. We also assessed the review authors' interpretation of results (as reported in the review conclusions), for high quality evidence and reports of statistically significant results.

2.~~.~~7~~6~~. Data synthesis and analysis

Descriptive statistics on year of publication of the update, as well as the time interval between the publication in the parent sample and the update were calculated. In addition, frequency of type of intervention and related outcome were calculated for the reviews that had been updated until the date of search. For studies that were updated, a change in the rating of evidence, if present, and its direction was recorded (downgrade, upgrade). Data accumulation for the review update was also recorded, based on number of studies/ participants included in the review's first listed outcome.

We reported actual proportions (n/N) as well as percentages of reviews reporting high, moderate, low or very low-quality evidence in the new sample of reviews. The quality of evidence according to GRADE in the new subset of reviews with updates was tabulated across the respective versions in the parent sample in a matched 4 x 4 table. We then

compared the difference in quality of evidence between the original and updated sample. We used the 2-sided exact signed-rank test to assess upgrades/downgrades between the original and updated reviews. We also performed a Stuart-Maxwell marginal homogeneity test. In addition, we performed assessments considering the presence of high-quality rating for any main outcome rather than just the first listed primary outcome.

For outcomes reported in the Summary of Findings table to be at the extremes (very low or high) of evidence quality, we reported the distribution of statistically significant results (P<0.05 or 95% confidence interval (CI) excluding the null), along with the reviewers' interpretation of the value of the intervention in clinical practice.

All statistical analyses were conducted with STATA software 15.1 (Stata Corporation, College Station, TX, USA) and R Software version 3.6.1 (R Foundation for Statistical Computing, Vienna, Austria).

~~*2.7. Outcomes*~~

~~The primary outcome was the change in quality of the evidence for the primary outcome in updated Cochrane Reviews compared with reviews published in an earlier (01/01/2013—30/06/2014) parent sample. The secondary outcomes were the proportion of reviews in the updated sample that have high, moderate, low, or very low-quality evidence. We also assessed the review authors' interpretation of results (as reported in the review conclusions), for high quality evidence and reports of statistically significant results.~~

2.8. Protocol Amendments

In the protocol, we planned a subgroup analyses by disease area, intervention type, and Cochrane Review Group. However, data for subgroups were deemed too sparse to allow for meaningful subgroup analyses.

3. Results

3.1. Search results

Of the 608 reviews in the original sample, 154 (25.3%) had been updated, and 151 of those presented information on GRADE quality of evidence for both initial and updated reviews so were retained for further assessment (Figure 1). The median year of the update was 2017 (interquartile range= 2, range: 2015 to 2020), with a median of 4 years (IQR= 2, range: 2 to 7 years) after the original review was published. Among the updated reviews, the original version with which it was compared (published in 2013-2014) was already an update of a previous version for 69 (45.7%) reviews.

Most reviews in the present samples of Cochrane updates pertained to pharmacological interventions (n=82; 54.4%), followed by behavioural (n=24; 15.9%) and surgical (n= 23; 15.2%) interventions, the use of medical devices (n=15; 9.9%), and diet- or exercise- related interventions (n=7; 4.6%). In most of the reviews, the primary outcome considered was classified as objective (127/151; 84.1%).

3.2. Quality of evidence in the entire updated (2020) sample

Within the 151 updated reviews, 15 (9.9%) had high quality evidence supporting the first listed primary outcome, 56 (37.1%) moderate, 47 (31.1%) low, and 33 (21.9%) very low. Compared with the original sample, there was a reduction in the proportion of reviews with high quality. However, this reduction was not statistically significant (see below). GRADE ranking comparison between the original and updated reviews are presented in Table 1, Table 2, and Figure 2.

Table 1. Summary of Review Quality from Updated and Original Samples

Year of review assessment	High N (%)	Moderate N (%)	Low N (%)	Very Low N (%)
2020	15 (9.9)	56 (37.1)	47 (31.1)	33 (21.9)
2013/14	82 (13.5)	187 (30.8)	193 (31.7)	146 (24)

Table 2. Change in quality of evidence across 151 reviews with updates for primary outcomes (the numbers below the diagonal are those which were upgraded, while those above were downgraded).

		GRADE quality of evidence in Updated Reviews (sample 2020)				
		High N (%)	Moderate N (%)	Low N (%)	Very Low N (%)	**Total**
GRADE quality of evidence in original sample (2013- 2014)	High N (%)	9 (60.0)	4 (7.1)	7 (14.9)	0 (0.0)	20 (13.2)
	Moderate N (%)	4 (26.7)	40 (71.4)	8 (17.0)	3 (9.1)	54 (35.8)
	Low N (%)	2 (13.3)	8 (14.3)	30 (63.8)	6 (18.2)	47 (31.1)
	Very Low N (%)	0 (0.0)	4 (7.2)	2 (4.3)	24 (72.7)	30 (19.9)
	Total	15 (100.0)	56 (100.0)	47 (100.0)	33 (100.0)	151 (100.0)

3.3. Change in quality of evidence

3.3.1. Change in quality of evidence for primary outcome

Most (103/151, 68.2%) of the updated reviews reported no change in the GRADE quality of evidence compared with the initial sample (blue diagonal in Table 1). Of the reviews with unchanged grading, 9 (8.7%) reported high-quality evidence, 40 (38.8%) had moderate, 30 (29.2%) low, and 24 (23.3%) very low quality of evidence. In 63 of the 103

updated reviews without a changed GRADE rating (61.2%), there was no additional data included in the updates, whereas in the remaining 35 reviews, more data had been added. In 5 reviews (4.9%) the update contained fewer primary studies than the original, but there was still no change in the GRADE rating. There was no statistical difference in the change in the quality of the evidence ratings (*P*= 0.30) between the original and updated reviews. The *P*-value for the marginal homogeneity test was 0.55.

A change in GRADE rating was reported in 48 of the 151 updated reviews. Twenty-eight of these (58.3%) were *downgraded*, mostly (24/28) to low or very low. Of first-listed primary outcomes initially recorded as having "high" quality evidence (n=15), 11 were downgraded to low (n=7) or moderate (n=4) quality of evidence. Twenty of the 48 reviews that had a changed GRADE involved an *upgrade*. Of those, 6 were upgraded to "high".

Thirty of the 48 trials (62.5%) that had a changed GRADE rating included additional data. Among these, 15 resulted in upgrades, and 15 in downgrades. In 16 (33.3%) the changed GRADE rating was not based on new data. In two updated reviews (4.2%), changes were based on fewer data for the primary outcome of interest; both resulted in upgrades. Finally, 16 out of 48 reviews with a change in GRADE rating, were based on the same included data (33.3%).

3.3.1. Change in quality of evidence for other outcomes (those that were not first listed non-primary)

Of the 151 updated reviews which did not present high quality of evidence for the first-listed primary outcome, 19 had other (non-primary, or primary but not first listed) outcomes that were ranked as high-quality. Ten of these involved primary outcomes. The overall quality of the evidence in the updates for any outcome was high in 34 out of 151 updated

reviews (22.5%). Again, we did not find a significant difference between the original and updated reviews for this comparison (P=0.72). The P-value by the marginal homogeneity test was P= 0.32.

3.4. Review authors' interpretations and statistical significance of results

Among extreme evidence quality ratings (very low and high), 8/33 (24.2 %) of those with very low quality and 10/15 (66.7 %) of those with high quality evidence had statistically significant results for at least one outcome in the updated sample. Across all 151 updated reviews, only 2 had high quality evidence, statistically significant results, and a favourable interpretation of the value of the intervention in clinical practice.

4. Discussion

4.1. Summary of findings

One-quarter of the reviews in our sample had been updated over the 6-7-year period. Of those, a third reported a change in GRADE ratings. There was no evidence of GRADE ratings being more likely to improve than worsen in these topics, with a weak trend towards worsening.

In keeping with a previous finding that 23% of Cochrane Reviews were out of date within two years, [11] our study may also show that Cochrane Reviews are not updated very frequently.[12] Specifically, we observed a median hiatus for publication of the updated review of 4 years among the reviews that were updated and most reviews were not even updated at all.

In some cases, downgrading of evidence quality was related to the new Risk of Bias assessment forming the basis for the GRADE framework. Risk of Bias assessments have become stricter in the new Cochrane Handbook and might have led to automatic downgrading due to items that had not been rated before or rated differently. This seems to be reflected in the fact that in approximately one-third of the reviews where the rating changed (16/48), there was no new data included in the review regarding the primary outcome of interest. Nevertheless, 81.3% (13/16) of the reviews with no new data reported worsening of evidence quality.

Another explanation for different GRADE ratings for updated reviews that had no new data is potential inconsistency in the way the way GRADE is applied. One study found variability in the way GRADE is applied leading to different conclusions about strength of evidence. [13] ~~Relatedly, a~~Another study found low agreement among systematic reviewers using the Cochrane Risk of Bias tool (which influences the GRADE rating). [14] This may~~ain~~ partially explain why two of the updated reviews whose evidence quality was upgraded were based on fewer studies than the original. The omitted studies also reduced imprecision or risk of bias. [15 16]

4.2. Limitations

The extent to which our findings are generalisable needs to be discussed. Our sample of reviews from 2013 and 2014 may not be representative of all medical evidence. It pertains to topics where either a new review was published at that time or it was deemed that an update was then indicated. Similarly, the reviews that were updated may not be representative of the original sample. Reviews which were not updated may have been less likely to require updating. If so, the proportion of changes in GRADE ratings we found may have been even

exaggerated. If we account also for this selection process, the results suggest that improvements in the quality of evidence in different medical topics are even more uncommon. Finally, we had a relatively small number of updated reviews, thus we could not meaningfully explore whether improvements in the quality of evidence are more or less likely in specific fields. However, no consistent patterns were observed for the very few reviews (n=6) where evidence was upgraded to high quality.

–In addition, our conclusions assumed that GRADE is sensitive enough to detect changes in evidence quality; this may not be necessarily~~solely~~ the case. GRADE only has four categories; if there ~~, and~~ were there additional categories, we may have detected a change in quality in a greater number of reviews. On top of that, GRADE assessments may suffer from inadequate interrater reliability, while evidence exists about training of review authors and/ or duplicate assessments on the use of GRADE, for an improved quality of the evidence evaluation approach. [17] On the other hand, a more sensitive evidence-rating tool could also be more likely to detect noise. More generally, our findings assumed that the GRADE ratings by the original review authors were reliable (and, more generally, that GRADE is reliable). ~~To overcome this limitation, a re-grading of the original and updated reviews would have to be undertaken by blinded reviewers.~~

Commented [1]: This is not a process that is easy to blind

4.3. Conclusion

Updating Cochrane systematic reviews does not change the fact that only a minority of outcomes for healthcare interventions are supported by high quality evidence. In spite of having additional data, most reviews were not updated over the time period of our assessment with the majority of updates not resulting in a change in the quality of the evidence. To avoid

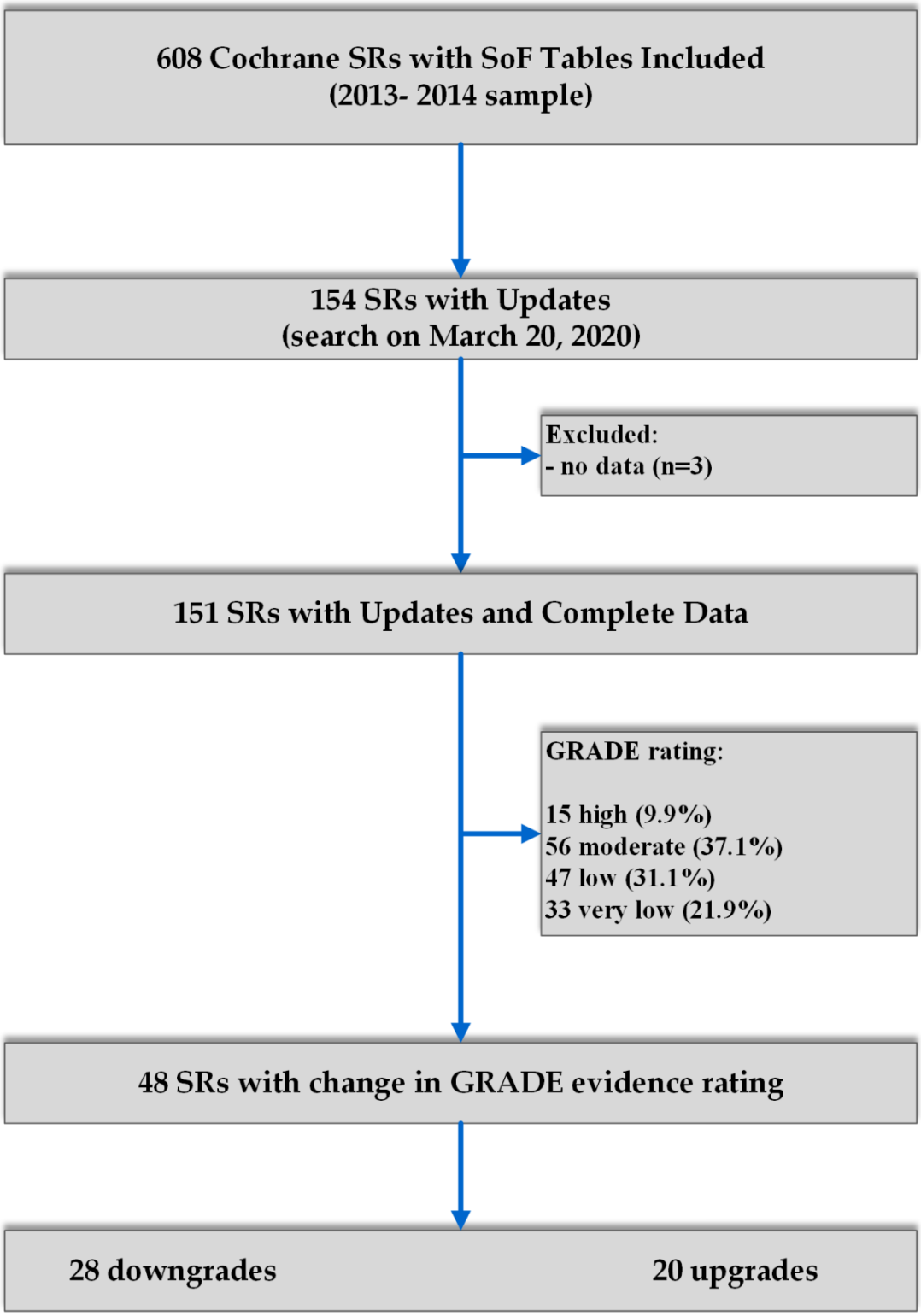
research waste, it should be investigated whether it is possible to decide in advance whether updating a review will result in a change in results. Effects of medical interventions supported by high quality evidence, statistically significant results, and favourable interpretations of the evidence by review authors remain rare.

References

1. Ezzo J, Bausell B, Moerman DE, et al. Reviewing the reviews. How strong is the evidence? How clear are the conclusions? *Int J Technol Assess Health Care* 2001;17(4):457-66. [published Online First: 2002/01/05]
2. Garrow JS. What to do about CAM: How much of orthodox medicine is evidence based? *BMJ* 2007;335(7627):951. doi: 10.1136/bmj.39388.393970.1F [published Online First: 2007/11/10]
3. El Dib RP, Atallah AN, Andriolo RB. Mapping the Cochrane evidence for decision making in health care. *J Eval Clin Pract* 2007;13(4):689-92. doi: 10.1111/j.1365-2753.2007.00886.x [published Online First: 2007/08/09]
4. Balshem H, Helfand M, Schunemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. *J Clin Epidemiol* 2011;64(4):401-6. doi: 10.1016/j.jclinepi.2010.07.015 [published Online First: 2011/01/07]
5. Fleming PS, Koletsi D, Ioannidis JP, et al. High quality of the evidence for medical and other health-related interventions was uncommon in Cochrane systematic reviews. *J Clin Epidemiol* 2016;78:34-42. doi: 10.1016/j.jclinepi.2016.03.012 [published Online First: 2016/04/02]
6. Chalmers I, Haynes B. Reporting, updating, and correcting systematic reviews of the effects of health care. *BMJ* 1994;309(6958):862-5. doi: 10.1136/bmj.309.6958.862 [published Online First: 1994/10/01]
7. Garritty C, Tsertsvadze A, Tricco AC, et al. Updating systematic reviews: an international survey. *PLoS One* 2010;5(4):e9914. doi: 10.1371/journal.pone.0009914 [published Online First: 2010/04/09]
8. Moher D, Tsertsvadze A. Systematic reviews: when is an update an update? *Lancet* 2006;367(9514):881-3. doi: 10.1016/S0140-6736(06)68358-X [published Online First: 2006/03/21]
9. French SD, McDonald S, McKenzie JE, et al. Investing in updating: how do conclusions change when Cochrane systematic reviews are updated? *BMC Med Res Methodol* 2005;5:33. doi: 10.1186/1471-2288-5-33 [published Online First: 2005/10/18]
10. Higgins JJ, Thomas JC, Chandler J, et al. Cochrane Handbook for Systematic Reviews of Interventions version 6.0. Version 6.0 ed. Chichester: The Cochrane Collaboration 2019.
11. Shojania KG, Sampson M, Ansari MT, et al. How quickly do systematic reviews go out of date? A survival analysis. *Ann Intern Med* 2007;147(4):224-33. doi: 10.7326/0003-4819-147-4-200708210-00179 [published Online First: 2007/07/20]
12. Higgins JJ, Green S. The Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 [updated March 2011] ed. Chichester: The Cochrane Collaboration 2011.
13. Berkman ND, Lohr KN, Morgan LC, et al. Interrater reliability of grading strength of evidence varies with the complexity of the evidence in systematic reviews. *J Clin Epidemiol* 2013;66(10):1105-17 e1. doi: 10.1016/j.jclinepi.2013.06.002 [published Online First: 2013/09/03]
14. Hartling L, Hamm MP, Milne A, et al. Testing the risk of bias tool showed low reliability between individual reviewers and across consensus assessments of reviewer pairs. *J Clin Epidemiol* 2013;66(9):973-81. doi: 10.1016/j.jclinepi.2012.07.005 [published Online First: 2012/09/18]
15. Boardman HM, Hartley L, Eisinga A, et al. Hormone therapy for preventing cardiovascular disease in post-menopausal women. *Cochrane Database Syst Rev* 2015(3):CD002229. doi: 10.1002/14651858.CD002229.pub4 [published Online First: 2015/03/11]
16. Hakoum MB, Kahale LA, Tsolakian IG, et al. Anticoagulation for the initial treatment of venous thromboembolism in people with cancer. *Cochrane Database Syst Rev* 2018;1:CD006649. doi: 10.1002/14651858.CD006649.pub7 [published Online First: 2018/01/25]

17. Mustafa RA, Santesso N, Brozek J, et al. The GRADE approach is reproducible in assessing the quality of evidence of quantitative evidence syntheses. *J Clin Epidemiol* 2013;66(7):736-42; quiz 42 e1-5. doi: 10.1016/j.jclinepi.2013.02.004 [published Online First: 2013/04/30]

Figure 1. Study selection and GRADE of evidence breakdown



What is new?

Key findings

- The quality of evidence (according to GRADE) supporting the main finding changes in about a quarter of updated Cochrane reviews.
- Upgrading of quality of evidence (according to GRADE) for the main outcome is not more common than downgrading of quality of evidence.

What this adds to what was known?

- Quality of evidence does not seem to improve overall with the addition of new evidence, at least within the timeframe assessed.

What is the implication and what should change now?

- Methods investigating when review updates are likely to change our confidence in the estimated outcome effect could inform decisions about whether to update reviews in order to save resources.
- The quality of evidence supporting most healthcare interventions remains low; higher quality evidence is required.

Has the Quality of Evidence for Medical Interventions Improved? A Meta-Epidemiological Study of Cochrane Reviews

Jeremy Howick, PhD[1], Despina Koletsi, DiplDS, Dr. med. dent[2*], Nikolaos Pandis[3], Padhraig S. Fleming, PhD[4], Martin Loef, PhD[5], Harald Walach, PhD[5,6], Stefan Schmidt, PhD[7], John P.A. Ioannidis, MD, DSc[8]

[1] Faculty of Philosophy, University of Oxford, Oxford OX2 6GG, United Kingdom

[2] Clinic of Orthodontics and Pediatric Dentistry, Center of Dental Medicine, University of Zurich, Switzerland *joint first author

[3] Department of Orthodontics and Dentofacial Orthopedics, School of Dental Medicine, Medical Faculty, University of Bern, Bern, Switzerland

[4] Institute of Dentistry, Queen Mary, University of London

[5] CHS-Institute, Berlin, Germany

[6] Poznan University of the Medical Sciences, Department of Pediatric Gastroenterlogy, Poznan, Poland

[7] Department of Psychosomatic Medicine and Psychotherapy, Medical Center, University of Freiburg

[8] Departments of Medicine, of Epidemiology and Population Health, of Biomedical Data Science, and of Statistics, and Meta-Research Innovation Center at Stanford (METRICS), Stanford University, CA, USA

Correspondence to: Jeremy Howick, Faculty of Philosophy, University of Oxford, Oxford OX2 6GG, +44 (0)7771925412, E-mail: jeremy.howick@philosophy.ox.ac.uk

Registration

Open Science Framework: Howick, J., Koletsi, D., Fleming, P., Schmidt, S., Loef, M., Walach, H., … Ioannidis, J. (2020, March 30). Has the Quality of Evidence for Medical Interventions Improved? Protocol for a Meta-Epidemiological Study. Retrieved from osf.io/bw7ky

Contributions

JH (guarantor) and JPAI conceived of the idea. JH wrote the first draft of the protocol. DK did the data extraction. JH, ML, PF, HW checked the extraction. DK and NP did the initial analysis. All authors interpreted the analyses, contributed to drafting the protocol and writing the manuscript.

Support

The writing of this protocol was not independently funded.

Declaration of interest

None of the authors have any conflicts of interests related to this paper.

Abstract

Background: A previous analysis of Cochrane Reviews published between January 1st, 2013 and June 30th, 2014 found that only 13.5% reported high quality evidence for the intervention according the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) system. 31.7% had low level, and 24% revealed very low level of evidence. Many of these reviews have been updated, and it is unknown whether the updated reviews report a change in the quality of evidence.

Objectives: To determine the change in quality of evidence in updates of Cochrane reviews that were initially published between 1st January 2013 and 30th June 2014.

Methods: We searched the Cochrane Database of Systematic Reviews on March 20th, 2020 to identify which of the reviews from the initial (2013/14) sample have been updated. Using the same methods to determine the quality of evidence in the previous analysis, we assessed the quality of evidence for the first listed primary outcomes in the updated reviews.

Results: Of the 608 reviews in the original sample, 154 had been updated with 151 presenting available data for both original and updated SRs (24.8%). The updated reviews included: 15 (9.9%) with high quality evidence, 56 (37.1%) with moderate, 47 (31.1%) with low, and 33 (21.9%) with very low-quality evidence. No change in the GRADE quality of evidence was found for most (103, 68.2%) of the updated reviews. Of the 48 reviews with a change in GRADE rating (58.3%) were downgraded, mostly to low or very low. The quality of evidence rating improved in 20 (41.7%), although only 6 reviews were promoted to high quality.

Conclusions: Updated systematic reviews continued to suggest that only a minority of outcomes for healthcare interventions are supported by high-quality evidence. The quality of the evidence did not consistently improve or worsen in updated reviews.

Keywords: Systematic review; evidence; Quality score; Meta-analysis

What is new?

Key findings

- The quality of evidence (according to GRADE) supporting the main finding changes in about a quarter of updated Cochrane reviews.
- Upgrading of quality of evidence (according to GRADE) for the main outcome is not more common than downgrading of quality of evidence.

What this adds to what was known?

- Quality of evidence does not seem to improve overall with the addition of new evidence, at least within the timeframe assessed.

What is the implication and what should change now?

- Methods investigating when review updates are likely to change our confidence in the estimated outcome effect could inform decisions about whether to update reviews in order to save resources.
- The quality of evidence supporting most healthcare interventions remains low; higher quality evidence is required.

1. Introduction

1.1. Rationale

Several meta-epidemiological studies have attempted to determine the proportion of healthcare interventions that are evidence-based. A 2001 estimate found that about a quarter (26.7%) of healthcare interventions whose effectiveness was reported in 160 Cochrane Reviews were considered effective, based on the interpretation of the review authors. [1] In 2007, Garrow claimed that 50% of healthcare treatments have good evidence to support them. [2] In the same year, El Dib *et al.* (2007) found that just 44% of a random selection of Cochrane Reviews evaluating interventions suggested that they were likely to be beneficial. [3]

Since these studies were published, the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) system has been introduced offering a less subjective way of ranking the quality of evidence. [4] An evaluation of all Cochrane Reviews published between January 1, 2013 and June 30, 2014 found that 13.5% of reviews were found to have high quality of evidence for the first listed primary outcome according to GRADE.[5] High quality evidence was more common in updated compared to new reviews and in association with pharmacologic than other types of interventions. Even when any outcomes (including but not limited to the first listed primary outcome) were considered, only 116/608 (19.1%) of the reviews reported at least one outcome with high quality of evidence.

Most researchers agree that it is important to update systematic reviews so that they reflect current knowledge, [6 7] to maximize patient benefits, and to avoid harm. [8] However, updated reviews frequently reveal no change in conclusions when compared with the original. According to French et al., only about 9% of updated Cochrane Reviews in 2002 presented a change in conclusion relative to their precursors from 1998. [9] However, the claim

that the updates did not overturn results from the original review was based on whether review authors stated there was a change in the conclusion of the updated review.

There is currently no consensus on the timing that would appropriately guide a review update and the Cochrane Collaboration's policy is to update reviews when evidence accumulates, based on the availability of new data that would have a meaningful impact on the findings and on the importance of the review question. [10] Previous reports have identified a median time required for an update of a systematic review of approximately 5.5 years. [11] It was therefore considered appropriate to assess whether reviews conducted back in 2013-2014 (Fleming et al., 2016) have been updated by early 2020, and if so, whether there are changes in the quality of the evidence based on GRADE. [5]

1.2. Objectives

The primary objective was to determine whether updates from a previous sample of systematic reviews resulted in a different quality evidence, as assessed by GRADE. The secondary objectives were to determine whether there is a difference in the change of quality of evidence across different interventions, outcomes, or Cochrane Review Groups.

2. Methods

2.1. Eligibility criteria

We included any Cochrane Review that was an update of a Cochrane Review published in the (01/01/2013—30/06/2014) parent sample of reviews which included a GRADE assessment.

2.2. Information sources

Cochrane Database of Systematic Reviews: https://www.cochranelibrary.com/cdsr/reviews.

2.3. Search strategy

We searched the Cochrane Database of Systematic Reviews to identify the reviews which had updates among those in the original sample. The most recent search was on March 20th, 2020.

2.4. Data sources and searches

One author (DK) retrieved the systematic reviews from the original (2013/14) sample and piloted the extraction form with one other author (JH). One author (DK) checked whether an update had been published and extracted data for the updated review. Other authors (JH, ML, PF, HW) were second extractors (all records were checked by two authors). All discrepancies were resolved by discussion.

2.5. Data items

Extracted information included: titles, corresponding author name and email, Cochrane Review Group, year of publication, country, study design, intervention (and intervention category), control and outcome. In relation to the GRADE Summary of Findings tables (SoF), we recorded for the first listed outcome the category of intervention (including surgical, pharmacologic, behavioural or medical treatments, and diet or exercise interventions). Interventions classified as: "behavioural" pertained to psychological treatment, psychotherapy, cognitive training, group therapy; "diet or exercise" interventions largely related to training exercise, physiotherapy, rehabilitation, dietary modification; "medical treatments" were summarized by electronic optical/ hearing aids, appliance/ device use for dental treatment, ultrasound or other radiography and medical interventions not related to surgical or pharmacologic approaches. We also recorded type of outcomes (objective, such as mortality or outcomes assessed with an instrument or pre-specified measurable criteria; or subjective) and overall GRADE ranking with reasons for downgrade or upgrade. In cases where multiple Summary of Findings tables within the same review existed for the primary outcome, we considered only the one listed first. In cases where no high-quality evidence was recorded for the first listed primary outcome, we documented whether any other outcome was rated as high and, if so, whether this was a primary (but not first listed) one.

We reported whether the Cochrane review authors concluded that the experimental intervention should be used in clinical or public health practice or not. This information was obtained from the conclusions section in the review abstract and the body of the review (subsections "implications for practice" and/ or "implications for research"), following the original strategy implemented in the parent study.[5] Examples of positive interpretations were: "Buprenorphine should be supported as a medication to use," and in the "Implications for

research or practice" section: "There does not appear to be any need for further randomized control trials of the relative efficacy of methadone compared with buprenorphine."

2.6. Outcomes

The primary outcome was the change in quality of the evidence for the primary outcome in updated Cochrane Reviews compared with reviews published in an earlier (01/01/2013—30/06/2014) parent sample. The secondary outcomes were the proportion of reviews in the updated sample that have high, moderate, low, or very low-quality evidence. We also assessed the review authors' interpretation of results (as reported in the review conclusions), for high quality evidence and reports of statistically significant results.

2.7. Data synthesis and analysis

Descriptive statistics on year of publication of the update, as well as the time interval between the publication in the parent sample and the update were calculated. In addition, frequency of type of intervention and related outcome were calculated for the reviews that had been updated until the date of search. For studies that were updated, a change in the rating of evidence, if present, and its direction was recorded (downgrade, upgrade). Data accumulation for the review update was also recorded, based on number of studies/ participants included in the review's first listed outcome.

We reported actual proportions (n/N) as well as percentages of reviews reporting high, moderate, low or very low-quality evidence in the new sample of reviews. The quality of evidence according to GRADE in the new subset of reviews with updates was tabulated across the respective versions in the parent sample in a matched 4 x 4 table. We then

compared the difference in quality of evidence between the original and updated sample. We used the 2-sided exact signed-rank test to assess upgrades/downgrades between the original and updated reviews. We also performed a Stuart-Maxwell marginal homogeneity test. In addition, we performed assessments considering the presence of high-quality rating for any main outcome rather than just the first listed primary outcome.

For outcomes reported in the Summary of Findings table to be at the extremes (very low or high) of evidence quality, we reported the distribution of statistically significant results ($P<0.05$ or 95% confidence interval (CI) excluding the null), along with the reviewers' interpretation of the value of the intervention in clinical practice.

All statistical analyses were conducted with STATA software 15.1 (Stata Corporation, College Station, TX, USA) and R Software version 3.6.1 (R Foundation for Statistical Computing, Vienna, Austria).

2.8. Protocol Amendments

In the protocol, we planned a subgroup analyses by disease area, intervention type, and Cochrane Review Group. However, data for subgroups were deemed too sparse to allow for meaningful subgroup analyses.

3. Results

3.1. Search results

Of the 608 reviews in the original sample, 154 (25.3%) had been updated, and 151 of those presented information on GRADE quality of evidence for both initial and updated reviews so were retained for further assessment (Figure 1). The median year of the update was 2017 (interquartile range= 2, range: 2015 to 2020), with a median of 4 years (IQR= 2, range: 2 to 7 years) after the original review was published. Among the updated reviews, the original version with which it was compared (published in 2013-2014) was already an update of a previous version for 69 (45.7%) reviews.

Most reviews in the present samples of Cochrane updates pertained to pharmacological interventions (n=82; 54.4%), followed by behavioural (n=24; 15.9%) and surgical (n= 23; 15.2%) interventions, the use of medical devices (n=15; 9.9%), and diet- or exercise- related interventions (n=7; 4.6%). In most of the reviews, the primary outcome considered was classified as objective (127/151; 84.1%).

3.2. Quality of evidence in the entire updated (2020) sample

Within the 151 updated reviews, 15 (9.9%) had high quality evidence supporting the first listed primary outcome, 56 (37.1%) moderate, 47 (31.1%) low, and 33 (21.9%) very low. Compared with the original sample, there was a reduction in the proportion of reviews with high quality. However, this reduction was not statistically significant (see below). GRADE ranking comparison between the original and updated reviews are presented in Table 1, Table 2, and Figure 2.

Table 1. Summary of Review Quality from Updated and Original Samples

Year of review assessment	High N (%)	Moderate N (%)	Low N (%)	Very Low N (%)

2020	15 (9.9)	56 (37.1)	47 (31.1)	33 (21.9)
2013/14	82 (13.5)	187 (30.8)	193 (31.7)	146 (24)

Table 2. Change in quality of evidence across 151 reviews with updates for primary outcomes (the numbers below the diagonal are those which were upgraded, while those above were downgraded).

		GRADE quality of evidence in Updated Reviews (sample 2020)				
		High N (%)	Moderate N (%)	Low N (%)	Very Low N (%)	**Total**
GRADE quality of evidence in original sample (2013- 2014)	High N (%)	9 (60.0)	4 (7.1)	7 (14.9)	0 (0.0)	20 (13.2)
	Moderate N (%)	4 (26.7)	40 (71.4)	8 (17.0)	3 (9.1)	54 (35.8)
	Low N (%)	2 (13.3)	8 (14.3)	30 (63.8)	6 (18.2)	47 (31.1)
	Very Low N (%)	0 (0.0)	4 (7.2)	2 (4.3)	24 (72.7)	30 (19.9)
	Total	15 (100.0)	56 (100.0)	47 (100.0)	33 (100.0)	151 (100.0)

3.3. Change in quality of evidence

3.3.1. Change in quality of evidence for primary outcome

Most (103/151, 68.2%) of the updated reviews reported no change in the GRADE quality of evidence compared with the initial sample (blue diagonal in Table 1). Of the reviews with unchanged grading, 9 (8.7%) reported high-quality evidence, 40 (38.8%) had moderate, 30 (29.2%) low, and 24 (23.3%) very low quality of evidence. In 63 of the 103

updated reviews without a changed GRADE rating (61.2%), there was no additional data included in the updates, whereas in the remaining 35 reviews, more data had been added. In 5 reviews (4.9%) the update contained fewer primary studies than the original, but there was still no change in the GRADE rating. There was no statistical difference in the change in the quality of the evidence ratings (P= 0.30) between the original and updated reviews. The P-value for the marginal homogeneity test was 0.55.

A change in GRADE rating was reported in 48 of the 151 updated reviews. Twenty-eight of these (58.3%) were *downgraded*, mostly (24/28) to low or very low. Of first-listed primary outcomes initially recorded as having "high" quality evidence (n=15), 11 were downgraded to low (n=7) or moderate (n=4) quality of evidence. Twenty of the 48 reviews that had a changed GRADE involved an *upgrade*. Of those, 6 were upgraded to "high".

Thirty of the 48 trials (62.5%) that had a changed GRADE rating included additional data. Among these, 15 resulted in upgrades, and 15 in downgrades. In 16 (33.3%) the changed GRADE rating was not based on new data. In two updated reviews (4.2%), changes were based on fewer data for the primary outcome of interest; both resulted in upgrades. Finally, 16 out of 48 reviews with a change in GRADE rating, were based on the same included data (33.3%).

3.3.1. Change in quality of evidence for other outcomes (those that were not first listed non-primary)

Of the 151 updated reviews which did not present high quality of evidence for the first-listed primary outcome, 19 had other (non-primary, or primary but not first listed) outcomes that were ranked as high-quality. Ten of these involved primary outcomes. The overall quality of the evidence in the updates for any outcome was high in 34 out of 151 updated

reviews (22.5%). Again, we did not find a significant difference between the original and updated reviews for this comparison (P=0.72). The P-value by the marginal homogeneity test was P= 0.32.

3.4. Review authors' interpretations and statistical significance of results

Among extreme evidence quality ratings (very low and high), 8/33 (24.2 %) of those with very low quality and 10/15 (66.7 %) of those with high quality evidence had statistically significant results for at least one outcome in the updated sample. Across all 151 updated reviews, only 2 had high quality evidence, statistically significant results, and a favourable interpretation of the value of the intervention in clinical practice.

4. Discussion

4.1. Summary of findings

One-quarter of the reviews in our sample had been updated over the 6-7-year period. Of those, a third reported a change in GRADE ratings. There was no evidence of GRADE ratings being more likely to improve than worsen in these topics, with a weak trend towards worsening.

In keeping with a previous finding that 23% of Cochrane Reviews were out of date within two years, [11] our study may also show that Cochrane Reviews are not updated very frequently.[12] Specifically, we observed a median hiatus for publication of the updated review of 4 years among the reviews that were updated and most reviews were not even updated at all.

In some cases, downgrading of evidence quality was related to the new Risk of Bias assessment forming the basis for the GRADE framework. Risk of Bias assessments have become stricter in the new Cochrane Handbook and might have led to automatic downgrading due to items that had not been rated before or rated differently. This seems to be reflected in the fact that in approximately one-third of the reviews where the rating changed (16/48), there was no new data included in the review regarding the primary outcome of interest. Nevertheless, 81.3% (13/16) of the reviews with no new data reported worsening of evidence quality.

Another explanation for different GRADE ratings for updated reviews that had no new data is potential inconsistency in the way the way GRADE is applied. One study found variability in the way GRADE is applied leading to different conclusions about strength of evidence. [13] Another study found low agreement among systematic reviewers using the Cochrane Risk of Bias tool (which influences the GRADE rating). [14] This may partially explain why two of the updated reviews whose evidence quality was upgraded were based on fewer studies than the original. The omitted studies also reduced imprecision or risk of bias. [15] [16]

4.2. Limitations

The extent to which our findings are generalisable needs to be discussed. Our sample of reviews from 2013 and 2014 may not be representative of all medical evidence. It pertains to topics where either a new review was published at that time or it was deemed that an update was then indicated. Similarly, the reviews that were updated may not be representative of the original sample. Reviews which were not updated may have been less likely to require updating. If so, the proportion of changes in GRADE ratings we found may have been even

exaggerated. If we account also for this selection process, the results suggest that improvements in the quality of evidence in different medical topics are even more uncommon. Finally, we had a relatively small number of updated reviews, thus we could not meaningfully explore whether improvements in the quality of evidence are more or less likely in specific fields. However, no consistent patterns were observed for the very few reviews (n=6) where evidence was upgraded to high quality.

In addition, our conclusions assumed that GRADE is sensitive enough to detect changes in evidence quality; this may not be necessarily the case. GRADE only has four categories; if there were there additional categories, we may have detected a change in quality in a greater number of reviews. On top of that, GRADE assessments may suffer from inadequate interrater reliability, while evidence exists about training of review authors and/ or duplicate assessments on the use of GRADE, for an improved quality of the evidence evaluation approach. [17] On the other hand, a more sensitive evidence-rating tool could also be more likely to detect noise. More generally, our findings assumed that the GRADE ratings by the original review authors were reliable (and, more generally, that GRADE is reliable).

4.3. Conclusion

Updating Cochrane systematic reviews does not change the fact that only a minority of outcomes for healthcare interventions are supported by high quality evidence. In spite of having additional data, most reviews were not updated over the time period of our assessment with the majority of updates not resulting in a change in the quality of the evidence. To avoid research waste, it should be investigated whether it is possible to decide in advance whether updating a review will result in a change in results. Effects of medical interventions

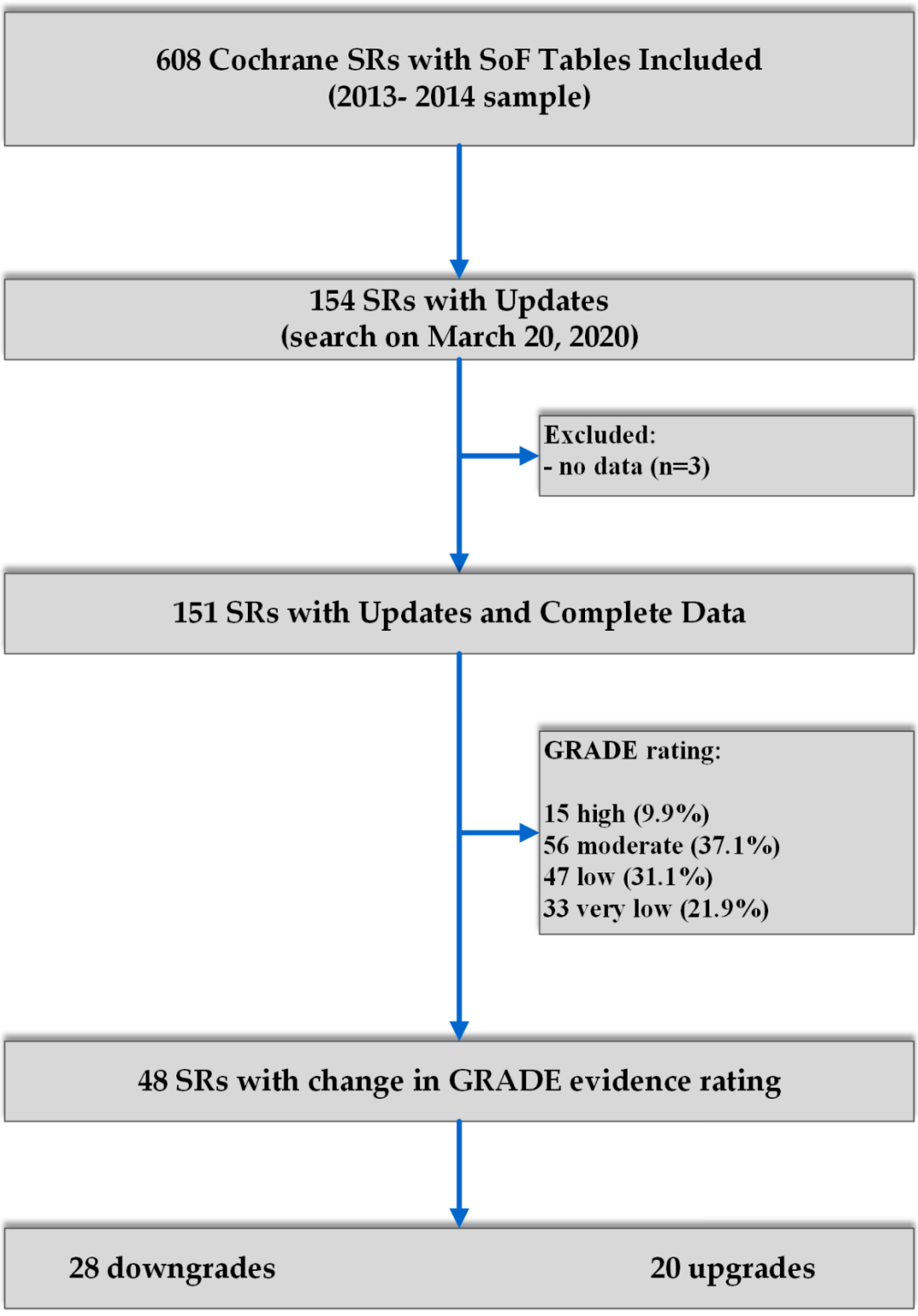
supported by high quality evidence, statistically significant results, and favourable interpretations of the evidence by review authors remain rare.

References

1. Ezzo J, Bausell B, Moerman DE, et al. Reviewing the reviews. How strong is the evidence? How clear are the conclusions? *Int J Technol Assess Health Care* 2001;17(4):457-66. [published Online First: 2002/01/05]
2. Garrow JS. What to do about CAM: How much of orthodox medicine is evidence based? *BMJ* 2007;335(7627):951. doi: 10.1136/bmj.39388.393970.1F [published Online First: 2007/11/10]
3. El Dib RP, Atallah AN, Andriolo RB. Mapping the Cochrane evidence for decision making in health care. *J Eval Clin Pract* 2007;13(4):689-92. doi: 10.1111/j.1365-2753.2007.00886.x [published Online First: 2007/08/09]
4. Balshem H, Helfand M, Schunemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. *J Clin Epidemiol* 2011;64(4):401-6. doi: 10.1016/j.jclinepi.2010.07.015 [published Online First: 2011/01/07]
5. Fleming PS, Koletsi D, Ioannidis JP, et al. High quality of the evidence for medical and other health-related interventions was uncommon in Cochrane systematic reviews. *J Clin Epidemiol* 2016;78:34-42. doi: 10.1016/j.jclinepi.2016.03.012 [published Online First: 2016/04/02]
6. Chalmers I, Haynes B. Reporting, updating, and correcting systematic reviews of the effects of health care. *BMJ* 1994;309(6958):862-5. doi: 10.1136/bmj.309.6958.862 [published Online First: 1994/10/01]
7. Garritty C, Tsertsvadze A, Tricco AC, et al. Updating systematic reviews: an international survey. *PLoS One* 2010;5(4):e9914. doi: 10.1371/journal.pone.0009914 [published Online First: 2010/04/09]
8. Moher D, Tsertsvadze A. Systematic reviews: when is an update an update? *Lancet* 2006;367(9514):881-3. doi: 10.1016/S0140-6736(06)68358-X [published Online First: 2006/03/21]
9. French SD, McDonald S, McKenzie JE, et al. Investing in updating: how do conclusions change when Cochrane systematic reviews are updated? *BMC Med Res Methodol* 2005;5:33. doi: 10.1186/1471-2288-5-33 [published Online First: 2005/10/18]
10. Higgins JJ, Thomas JC, Chandler J, et al. Cochrane Handbook for Systematic Reviews of Interventions version 6.0. Version 6.0 ed. Chichester: The Cochrane Collaboration 2019.
11. Shojania KG, Sampson M, Ansari MT, et al. How quickly do systematic reviews go out of date? A survival analysis. *Ann Intern Med* 2007;147(4):224-33. doi: 10.7326/0003-4819-147-4-200708210-00179 [published Online First: 2007/07/20]
12. Higgins JJ, Green S. The Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 [updated March 2011] ed. Chichester: The Cochrane Collaboration 2011.
13. Berkman ND, Lohr KN, Morgan LC, et al. Interrater reliability of grading strength of evidence varies with the complexity of the evidence in systematic reviews. *J Clin Epidemiol* 2013;66(10):1105-17 e1. doi: 10.1016/j.jclinepi.2013.06.002 [published Online First: 2013/09/03]
14. Hartling L, Hamm MP, Milne A, et al. Testing the risk of bias tool showed low reliability between individual reviewers and across consensus assessments of reviewer pairs. *J Clin Epidemiol* 2013;66(9):973-81. doi: 10.1016/j.jclinepi.2012.07.005 [published Online First: 2012/09/18]
15. Boardman HM, Hartley L, Eisinga A, et al. Hormone therapy for preventing cardiovascular disease in post-menopausal women. *Cochrane Database Syst Rev* 2015(3):CD002229. doi: 10.1002/14651858.CD002229.pub4 [published Online First: 2015/03/11]
16. Hakoum MB, Kahale LA, Tsolakian IG, et al. Anticoagulation for the initial treatment of venous thromboembolism in people with cancer. *Cochrane Database Syst Rev* 2018;1:CD006649. doi: 10.1002/14651858.CD006649.pub7 [published Online First: 2018/01/25]
17. Mustafa RA, Santesso N, Brozek J, et al. The GRADE approach is reproducible in assessing the quality of evidence of quantitative evidence syntheses. *J Clin Epidemiol*

2013;66(7):736-42; quiz 42 e1-5. doi: 10.1016/j.jclinepi.2013.02.004 [published Online First: 2013/04/30]

Figure 1. Study selection and GRADE of evidence breakdown



Declaration of interest

None of the authors have any conflicts of interests related to this paper.

Contributions

JH (guarantor) and JPAI conceived of the idea. JH wrote the first draft of the protocol. DK did the data extraction. JH, ML, PF, HW checked the extraction. DK and NP did the initial analysis. All authors interpreted the analyses, contributed to drafting the protocol and writing the manuscript.