Journal of Assisted Reproduction and Genetics
https://doi.org/10.1007/s10815-019-01663-y

COMMENTARY

Quality of evidence matters: is it well reported and interpreted in infertility journals?

Demian Glujovsky[1,2,3] · Carlos E. Sueldo[2,4] · Ariel Bardach[1] · María del Pilar Valanzasca[5] · Daniel Comandé[1] · Agustín Ciapponi[1,4]

Received: 13 August 2019 / Accepted: 12 December 2019

Abstract

Purpose To evaluate if the authors of published systematic reviews (SRs) reported the level of quality of evidence (QoE) in the top 5 impact factor infertility journals and to analyze if they used an appropriate wording to describe it.

Methods This is a cross-sectional study. We searched in PubMed for SRs published in 2017 in the five infertility journals with the highest impact factor. We analyzed the proportion of SRs published in the top 5 impact factor infertility journals that reported the SRs' QoE, and the proportion of those SRs in which authors used consistent wording to describe QoE and magnitude of effect.

Results The QoE was reported in only 21.4% of the 42 included SRs and in less than 10% of the abstracts. Although we did not find important differences in the report of QoE of those that showed statistically significant differences or not, *p* value was associated with the wording chosen by the authors. We found inconsistent reporting of the size the effect estimate in 54.8% (23/42) and in the level of QoE in 92.9% (39/42). Whereas the effect size was more consistently expressed in studies with statistically significant findings, QoE was better expressed in those cases in which the *p* value was over 0.05.

Conclusion We found that in 2017, less than 25% of the authors reported the overall QoE when publishing SRs. Authors focused more on statistical significance as a binary concept than on methodological limitations like study design, imprecision, indirectness, inconsistency, and publication bias. Authors should make efforts to report the QoE and interpret results accordingly.

Keywords Quality of evidence · Systematic reviews · Magnitude of the effect

Introduction

Systematic reviews (SRs) aim to critically analyze and integrate all the information from primary research on a specific topic. Meta-analysis (MA) is the statistical procedure for combining data from multiple studies, identified through SRs, to obtain a combined effect. The strength of the evidence that they provide depends mainly on the quality of the primary studies identified.

Traditionally, the hierarchy of evidence regarding therapy or prevention issues has been represented as a pyramid, placing the strongest designs such as randomized controlled trials (RCTs) and SRs/MAs at the top. More recently, a new evidence-based medicine pyramid has been proposed, removing the SRs from the top of the pyramid and using them as lenses through which other types of studies should be seen [1]. Despite the increasing use and popularity of SRs, it is crucial to assess the certainty or quality of evidence (QoE) they provide, since the design of the primary study itself does not necessarily guarantee a high QoE. The Grading of Recommendations Assessment, Development and Evaluations (GRADE) approach is the soundest system for rating the certainty of a body of evidence in systematic reviews and other evidence syntheses. Although GRADE

✉ Demian Glujovsky
glujovsky@cegyr.com

[1] Argentine Cochrane Centre, Institute for Clinical Effectiveness and Health Policy (IECS), Center for Research in Epidemiology and Public Health, National Scientific and Technical Research Council (CONICET), Ravignani 2024, C1414CPV Buenos Aires, Argentina

[2] Center for Studies in Genetics and Reproduction (CEGYR), Buenos Aires, Argentina

[3] Centro de Estudios en Genética y Reproducción (CEGYR), Viamonte 1432, C1055ABB Buenos Aires, Argentina

[4] Obstetrics and Gynecology Department, University of California San Francisco-Fresno, 155 N Fresno St Suite 233, Fresno, CA 93701, USA

[5] Aventura, FL, USA

Published online: 23 December 2019

currently prefers the use of the word "certainty" [2], we used the word "quality" throughout this paper because it is probably better known to the target audience of this manuscript. GRADE offers a transparent and structured process for assessing the evidence, developing and presenting evidence summaries, and making recommendations [3]. GRADE provides explicit criteria for rating the QoE that includes study design, imprecision (wide confidence interval), inconsistency (unexplained heterogeneity of results), indirectness (differences between the target and the actual population, intervention, and outcomes of interest), and publication bias (a systematic under/overestimation of the underlying effect due to the selective publication of studies). These factors can downgrade the QoE. As a distinctive approach, the QoE is rated for each outcome considering the whole body of evidence, and it is classified into four levels (high, moderate, low, and very low) based on the confidence on the SR estimates [3]. The QoE and the magnitude of relative and absolute effects for each important outcome are presented together in standardized Summary of Findings tables, including explanations in support of factors that affect the QoE grading for each outcome.

Except for the Cochrane reviews, which report QoE as established in its manual [4], QoE reporting is infrequent in non-Cochrane systematic reviews. Even less frequent is that the level of evidence and the magnitude of the effect are interpreted using an appropriate wording. We have already described different types of methodological flaws in RCTs published in top 5 impact factor infertility journals [5], but we are describing the findings in SRs in the present work. A common mistake is to confuse "no evidence of an effect" with "evidence of no effect." When the confidence intervals are too wide, crossing the line of "no effect," claiming that the intervention has "no effect" or that it is "not different" from the comparative intervention could be false. When the effect estimate for an outcome is beneficial, but confidence interval is too wide, authors often describe the effect as "promising." However, in a similar situation, when the effect estimate is considered harmful but the confidence interval includes the possibility of no effect, authors frequently report "no effect" [6–9]. Another misleading practice is to frame the conclusion in wishful terms. For example, if a confidence interval is compatible with both a true beneficial effect and a true harmful effect, both possibilities should be mentioned. Quality of the evidence is very important to understand the extent to which the results of a research are reliable. However, QoE is rarely evaluated or elaborated in the discussion section. When QoE was evaluated and reported, the effect of the *p* value still weighs more than the QoE to draw conclusions [10].

Cochrane groups have proposed narrative statements for drawing conclusions based on the magnitude of the effect estimate and the certainty of the evidence of the meta-analyses and provided a good framework for the analysis in the last Cochrane Handbook version [7, 8, 11]. Although there is not a unique valid wording, authors should make an effort to adapt their words to these two important concepts.

The aim of this study was to describe if the authors of SRs of interventions in the field of infertility assessed and reported the QoE of primary research in systematic reviews. In addition, we analyzed if the level of QoE and magnitude of the effect were consistent with the wording used by the authors in results and conclusions.

Material and methods

We performed a cross-sectional study. As we did in some of our previous studies in which quality of research was evaluated [5, 12] by our group, we selected the five journals focused on human reproduction with the highest impact factors, according to the most recent published list that is the 2018 Impact Factor (from the Institute for Scientific Information) and the H-index (from SciMagO) [13]. They were *Human Reproduction Update* (IF 12.878), *Human Reproduction* (IF 5.506), *Fertility and Sterility* (IF 5.411), *Reproduction* (Cambridge, England; IF 3.125), and *Reproductive Biomedicine Online* (IF 2.930). Impact factor measures if science journals' publications are cited, and importance of a journal could be estimated using this parameter [14]. Impact factor is used by many authors to rank journals [15]. Firstly, we performed a search in September 2018 in PubMed, identifying the potential SRs with meta-analysis (limits, type of article: meta-analysis) published in the year 2017, the most recent complete year at the moment when the search was done. Secondly, we screened the studies by title and abstract to include only those that were systematic reviews of intervention and where the main subject of study was infertility. We excluded those studies researching topics like contraception, menopause, any other study not done in humans, or any study that was not a systematic review. After that, two reviewers (DG, AC, PV, AB) randomly selected, extracted, and independently assessed each full text for final eligibility. Discrepancies were resolved by consensus. We considered a systematic review as per the classification of the Database of Abstracts of Reviews of Effect (DARE) [16]. Once we retrieved the full text of the systematic reviews, we checked if the authors evaluated QoE and if they used any specific tool. We analyzed if they classified the published evidence in the full text and also if they did it in the abstract. When authors did not evaluate the QoE, we used GRADE [3] to analyze it. We described how often a tool for QoE was used by the authors and the level of the QoE published in the selected systematic reviews. We also analyzed if using a tool for QoE was associated with the *p* value or not, and if the *p* value was associated with the level of the evidence or not. Finally, we analyzed if the authors made any effort to adapt the wording used in the

abstract to the QoE and the magnitude of the described estimated effect. For this purpose, we compared the wording used with the wording proposed both by Cochrane and Support [6–8].

We used Covidence, a web-based software to facilitate these phases of classification [17]. We conducted a descriptive analysis of all evaluated articles, and data was analyzed using STATA version 11.1 (StataCorp). Pearson's chi-square statistics were used for determining the degree of independence between categorical variables. All the authors stated having no conflicts of interest because they do not participate in any editorial board of reproductive medicine journals.

Results

General description

We screened 178 studies by title and abstract, and we selected 50 studies to be screened by full text. The rest were excluded due to the methods or the topic. Of those, eight were excluded because they did not comply with the requisites needed to be considered systematic review. The QoE of the included SRs was evaluated by the authors of those publications in only nine out of the 42 included studies (21.4%). In those in which the quality was analyzed, most authors used GRADE (77.8%), while the rest used other tools. However, when reporting results in the abstract, the level of the QoE was mentioned in less than half (4/9; 44.4%) of those SRs in which it was assessed, meaning that QoE was stated in the abstract in less than 10% (4/42) of all the included SRs.

We found no important differences between SRs in which QoE was evaluated and those in which it was not. When we analyzed the primary outcomes, we found that in 23 studies (54.8%), the difference between the intervention and its comparator was reported as statistically significant. We did not find important differences in the report of QoE among those that showed statistically significant differences vs those that did not (8.7% vs 10.5%, $p = 0.84$).

QoE was not assessed in 33 of 42 SRs (78.5%, 95%CI 63.1–89.7), and when we analyzed the QoE when authors did not, we did not find any bias that suggests that authors decided to analyze the QoE according to the level of evidence encountered (see Table 1).

Wording used to describe the results

The reported magnitude of the effect was not expressed with consistent wording in 23 out of 42 (54.8%, 95%CI 38.7–70.1) SRs, while the level of QoE was not expressed with consistent wording in 39 out of 42 (92.9%, 95%CI 80.5–98.5) SRs.

Reporting of statistically significant differences in the primary outcome was associated with the consistency of the wording used by the authors in the abstract. When we analyzed the wording used to state the magnitude of the effect, we found that it was more consistently expressed in those cases where the difference was statistically significant whereas in cases where the p value was over 0.05, the magnitude of the effect was described in fewer instances (65.2% vs 21.1%; $p = 0.004$). However, when we analyzed the wording used to state the level of evidence, we found that the level of QoE was better expressed in those cases in which the difference was not statistically significant, while in those cases where the p value was below 0.05, QoE was always poorly described (0% vs 15.8%; $p = 0.048$). Table 2 reports the wording used by authors in those cases in which the magnitude of the effect was not shown by the wording used in the abstract.

Discussion

In this study, we found that in 2017, in the top 5 impact factor infertility journals, less than 25% of authors reported the overall QoE when publishing systematic reviews, and even less than one in 10 reported QoE in the abstract. It was not clear if it was not reported because QoE was not considered important, or because they did not know how to apply it or interpret it. In fact, when QoE was reported, authors did not usually choose an appropriate wording. We should emphasize that not only authors missed to evaluate the QoE but also peer-reviewers missed to note the QoE absence, or consider that its absence was important. Although the absence of

Table 1 Quality of evidence (QoE) level reported or not by the authors

	Reported by the authors $N = 9$	Not reported by the authors* $N = 33$
High evidence	0%	9.1% (3/33)
Moderate evidence	33.3% (3/9)	33.3% (11/33)
Low evidence	22.2% (2/9)	18.2% (6/33)
Very low evidence	44.4% (4/9)	39.4% (13/33)
Total	100% (9/9)	100% (33/33)

P = NS

*QoE was analyzed by our team

Table 2 Three illustrating examples of the wording used by authors that did not help to understand the magnitude of the effect

QoE	Magnitude	Wording used by the authors	Reason for potential misleading	Alternative option
M	RR 1.25 (0.96–1.63)	There was no statistically significant difference in terms of pregnancy rate	Although this expression is not incorrect, it does not state anything about the magnitude of effect size and the QoE	Based on the central estimation, the intervention probably increases the pregnancy rate. However, the confidence interval limits are compatible with both increase and decrease
L	OR 0.62 (0.22–1.7)	No significant differences.	The central estimation is far from the non-effect line. That is not shown in the wording used. And, the quality of the evidence is low, which is not shown either.	Based on the central estimation, the evidence suggests that AAS may reduce the hypertensive complications. However, the confidence interval limits are compatible with both increase and decrease
VL	RR 1.48 (1.09–2.02)	Blastocyst stage transfer was associated with increased risks of preterm birth	As QoE is very low, stating any association could be misleading	The evidence is very uncertain about the effect of blastocyst stage transfer on preterm birth.

assessment of QoE was not associated with any specific outcome result or any specific characteristic found in the systematic reviews (which would have meant a bias), it could limit the interpretation of the reliability of the results.

A second important finding was that authors usually interpreted the level of the QoE and the description of the magnitude of the effect very poorly; the wording used by the authors in results and discussion was more associated with the *p* value than with the overall quality of evidence. In other words, *p* value was given more weight than the level of the QoE and the magnitude of effect in the wording used by the authors.

In those cases in which the association between the intervention and the outcome was statistically significant, the wording used to describe the magnitude of the effect was consistent with the magnitude itself. However, authors were not accurate enough to describe the magnitude in those cases in which the *p* value was higher than 0.05. On the contrary, authors were more accurate to describe the QoE when the association was not statistically significant. In those cases in which the *p* value was low, but the QoE was not good, the consistency was not good either. In summary, authors were less accurate to describe the magnitude in those cases in which the *p* value was higher than the cutoff point and were less accurate to describe the QoE when *p* values were lower.

Recently, hundreds of scientists rose up against the reporting of statistical significance as a binary concept. The false belief that only reaching statistical significance indicates that a result is "real" has led scientists and journal editors to legitimize those results, and in this way, distort the literature [9].

A strength of our study is that we used strict criteria to classify systematic reviews, but if we would have analyzed all the studies that were labeled as systematic reviews, the proportion of QoE reporting probably would have been much lower. On the other hand, as a limitation, we should say that QoE classification involves subjective judgments and there could be some inter-reviewers' variability, even when all the reviewers are experts in evidence research, and even if they used GRADE which is probably the most popular and accepted tool for this purpose. In order to address this issue, two independent reviewers analyzed each of the included studies. Another weakness of our study is that, although Cochrane and Support [6–8] suggest some wording to gain consistency, this wording is not widely accepted or known by all authors and editorial boards. Therefore, it should not be considered totally wrong when authors do not use it. Nevertheless, we did not classify the studies as consistent or not, according to Cochrane or Support wording proposals, but we analyzed if authors made any effort to use specific wording to avoid readers to be misled by inconsistencies, between effect size, statistics, and quality of evidence. Finally, as we limited our assessment to systematic reviews that were published in the most cited

infertility journals, we cannot generalize these results as this sample could not adequately represent the whole body of evidence coming from all the infertility journals.

When looking for some other studies evaluating the importance of the interpretation of results, we found a study published by McGrath et al. that analyzed 112 systematic reviews and stated that 72% of them contained at least one form of overinterpretation in the abstract and 69% in the full text [18]. These authors warn about the risk of making erroneous clinical decisions and recommendations when not being accurate with the interpretation of the results. On the other hand, both Lumbreras et al. and Ochodo et al. published two papers in which they found that overinterpretation of clinical applicability in molecular diagnostic research is higher in high impact factor journals, which they discuss as a concerning feature that should be improved [19, 20].

The present study shows that authors pay more attention to discuss if the difference encountered was by chance or not, and less attention to limitations in the study design, imprecision, indirectness, inconsistency among the primary studies, and publication bias. Even size effect was not considered when retrieving conclusions. Whenever authors submit a study for publication, QoE is very important to understand the extent of the conclusions that we can arrive at. When analyzing systematic reviews, reviewers can use AMSTAR 2 [21] and PRISMA [22] to evaluate how the review was undertaken and how it was reported, respectively. However, authors should present their results, including the overall evaluation of QoE coming from primary studies, which help readers and policy makers to better understand the importance of the findings published in that systematic review. Besides, once the evaluation of QoE has been done, authors should make efforts to interpret those results in the context of that evaluation. It is important not only to mention the level of QoE but also to use an appropriate wording to describe the results in order to avoid a potential misleading interpretation by the readers. Of course, this is not only a recommendation for authors, it is also for editorial boards and referees. Both those who submit the reports and those who review them should stress the need for analyzing and reporting the QoE in this type of research, thereby improving the quality of the publications in infertility journals.

Systematic reviews are considered one of the study designs that provide the best QoE. But, its QoE depends mainly on the quality of each included primary study and the result that every one of them provide. If authors do not make any effort to analyze QoE, readers could misinterpret the results. Systematic reviews per se do not guarantee high QoE, as it depends partially on the included primary studies. In cases when QoE is low or very low, readers should clearly understand that evidence is uncertain. When authors analyze QoE but do not use a consistent wording to describe the results and conclusions, they could be misleading the readers.

Our study highlights an opportunity for improvement. In the future, in order to optimize the QoE reporting in SRs, authors could be more consistent when retrieving conclusions. There are some initiatives to use narrative statements that are becoming more popular, such as the one described in the Cochrane Handbook [6]. On the other hand, editorial boards could have stricter standards for review, including some options of statements in the instructions for authors.

There are still many unanswered questions: about which is the best tool to evaluate QoE? Which is the best wording set to help readers to better interpret the results? How to deal with confidence intervals that are very close to the non-effect line? What is considered large effect and small effect? How is the best way to state minimal important difference (MID) of a result? These are just some of the important questions that have to be addressed in future research.

Conclusions

We found that just one in five SRs of interventions in the field of infertility assessed and reported the QoE of primary research. In addition, we found that less than half of the SRs used a wording that was consistent with the effect magnitude and just one in 10 used a wording consistent with the level of QoE. Finally, when expressing results, authors of SRs paid more attention to the statistical significance, as a binary concept, rather than to the methodological limitations like study design, imprecision, indirectness, inconsistency, and publication bias.

The findings in our study should be important for authors and editorial boards at infertility journals. A more consistent wording should help in giving a more accurate message to readers and policy makers.

Acknowledgments The authors would like to thank Heike Thiel for the review of the paper.

References

1. Murad MH, Asi N, Alsawas M, Alahdab F. New evidence pyramid. Evid Based Med. 2016;21(4):125–7. https://doi.org/10.1136/ebmed-2016-110401.
2. Hultcrantz M, Rind D, Akl EA, Treweek S, Mustafa RA, Iorio A, et al. The GRADE rWorking Group clarifies the construct of certainty of evidence. J Clin Epidemiol. 2017;87:4–13. https://doi.org/10.1016/j.jclinepi.2017.05.006.
3. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011;64(4):383–94. https://doi.org/10.1016/j.jclinepi.2010.04.026.
4. Higgins J, Green S, (editors). Cochrane handbook for systematic reviews of interventions version 5.1.0 [updated March 2011]. 2011.
5. Glujovsky D, Sueldo CE, Borghi C, Nicotra P, Andreucci S, Ciapponi A. Misleading reporting and interpretation of results in

major infertility journals. Fertil Steril. 2016. https://doi.org/10.1016/j.fertnstert.2015.12.134.
6. Schünemann HJ, Vist GE, Higgins JPT, Santesso N, Deeks JJ, Glasziou P et al. Chapter 15: Interpreting results and drawing conclusions. In: Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page M et al. editors. Cochrane handbook for systematic reviews of interventions version 6 [updated September 2018]: Cochrane. 2018.
7. Cochrane Effective Practice and Organisation of Care (EPOC). Reporting the effects of an intervention in EPOC reviews. EPOC Resources for review authors (Version: 24 August 2017). 2017. http://epoc.cochrane.org/sites/epoc.cochrane.org/files/public/uploads/Resources-for-authors2017/how_to_report_the_effects_of_an_intervention.pdf. Accessed 19/12/2017.
8. Reporting results in CKT reviews (using material adapted from EPOC and CCCR). 2017. http://kidneyandtransplant.cochrane.org/sites/kidneyandtransplant.cochrane.org/files/public/uploads/Resources/reporting_results_in_ckt_reviews_2017.pdf. Accessed 19/12/2017.
9. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature. 2019;567(7748):305–7. https://doi.org/10.1038/d41586-019-00857-9.
10. Ciapponi A, Glujovsky D, Comande D, Bardach A, editors. Do Cochrane systematic reviews report results integrating certainty of evidence and effect size? 25th Cochrane Colloquium. Scotland: Edinburgh; 2018.
11. Higgins J, Thomas J, Cumpston M, Chandler J, Li T, Page M, et al. Cochrane handbook for systematic reviews of interventions version 6: DRAFT. 2018.
12. Glujovsky D, Riestra B, Coscia A, Boggino C, Comande D, Ciapponi A. Assessment of research quality in major infertility journals. Fertil Steril. 2012;98(6):1539–43. https://doi.org/10.1016/j.fertnstert.2012.08.018.
13. Scimago Journal & Country Rank [Portal]. SCImago, (n.d.). SJR. 2019. https://www.scimagojr.com/journalrank.php. Accessed 01/15/2019.
14. CiteFactor. 2019. https://www.citefactor.org/. Accessed 01/15/2019.
15. Reveiz L, Cortes-Jofre M, Asenjo Lobos C, Nicita G, Ciapponi A, Garcia-Dieguez M, et al. Influence of trial registration on reporting quality of randomized trials: study from highest ranked journals. J Clin Epidemiol. 2010;63(11):1216–22. https://doi.org/10.1016/j.jclinepi.2010.01.013.
16. Database of Abstracts of Reviews of Effects (DARE): quality-assessed reviews. York (UK): Centre for Reviews and Dissemination (UK). 2019. https://www.ncbi.nlm.nih.gov/books/NBK285222/. Accessed 01/15/2019.
17. Covidence systematic review software. Veritas Health Innovation: Melbourne. www.covidence.org.
18. McGrath TA, McInnes MDF, van Es N, Leeflang MMG, Korevaar DA, Bossuyt PMM. Overinterpretation of research findings: evidence of "spin" in systematic reviews of diagnostic accuracy studies. Clin Chem. 2017;63(8):1353–62. https://doi.org/10.1373/clinchem.2017.271544.
19. Lumbreras B, Parker LA, Porta M, Pollan M, Ioannidis JP, Hernandez-Aguado I. Overinterpretation of clinical applicability in molecular diagnostic research. Clin Chem. 2009;55(4):786–94. https://doi.org/10.1373/clinchem.2008.121517.
20. Ochodo EA, de Haan MC, Reitsma JB, Hooft L, Bossuyt PM, Leeflang MM. Overinterpretation and misreporting of diagnostic accuracy studies: evidence of "spin". Radiology. 2013;267(2):581–8. https://doi.org/10.1148/radiol.12120527.
21. 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. https://doi.org/10.1136/bmj.j4008.
22. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 2009;6(7):e1000097. https://doi.org/10.1371/journal.pmed.1000097.