Vali et al. Systematic Reviews (2021) 10:191
https://doi.org/10.1186/s13643-021-01744-z

Systematic Reviews

METHODOLOGY Open Access

Application of weighting methods for presenting risk-of-bias assessments in systematic reviews of diagnostic test accuracy studies

Yasaman Vali*, Mariska M. G. Leeflang and Patrick M. M. Bossuyt

Abstract

Background: An assessment of the validity of individual diagnostic accuracy studies in systematic reviews is necessary to guide the analysis and the interpretation of results. Such an assessment is performed for each included study and typically reported at the study level. As studies may differ in sample size and disease prevalence, with larger studies contributing more to the meta-analysis, such a study-level report does not always reflect the risk of bias in the total body of evidence. We aimed to develop improved methods of presenting the risk of bias in the available evidence on diagnostic accuracy of medical tests in systematic reviews, reflecting the relative contribution of the study to the body of evidence in the review.

Methods: We applied alternative methods to represent evaluations with the Quality Assessment of Diagnostic Accuracy Studies tool (QUADAS-2), weighting studies according to their relative contribution to the total sample size or their relative effective sample size. We used these methods in four existing systematic reviews of diagnostic accuracy studies, including 9, 13, 22, and 32 studies, respectively.

Results: The risk-of-bias summaries for each domain of the QUADAS-2 checklist changed in all four sets of studies after replacing unit weights for the studies with relative sample sizes or with the relative effective sample size. As an example, the risk of bias was high in the patient selection domain in 31% of the studies in one review, unclear in 23% and low in 46% of studies. Weighting studies according to the relative sample size changed the corresponding proportions to 4%, 4%, and 92%, respectively. The difference between the two weighting methods was small and more noticeable when the reviews included a smaller number of studies with wider range of sample size.

Conclusions: We present an alternative way of presenting the results of risk-of-bias assessments in systematic reviews of diagnostic accuracy studies. Weighting studies according to their relative sample size or their relative effective sample size can be used as more informative summaries of the risk of bias in the total body of available evidence.

Systematic review registrations: Not applicable

Keywords: Risk-of-bias assessment, Quality appraisal, Diagnostic accuracy studies, Systematic reviews

* Correspondence: y.vali@amsterdamumc.nl
Department of Epidemiology and Data Science, Amsterdam UMC, University of Amsterdam, Meibergdreef 9, 1105AZ, Amsterdam, The Netherlands

© 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.

Background

Systematic reviews are important tools in evidence synthesis, particularly for combining the results of multiple primary studies which may have conflicting results [1–3]. The credibility of a systematic review depends heavily on the methodological quality of included studies, which impacts the credibility of the findings and the strength of the final conclusions of the review [4]. It is therefore essential that reviewers thoroughly assess the validity of included studies, to appraise the certainty of the evidence in the review and to draw conclusions confidently.

Assessing the risk of bias in primary studies is a fundamental component of systematic reviews. It helps to establish transparency of evidence synthesis results, supports the interpretation of findings and explanations of heterogeneity. Existing guidelines, such as the Cochrane handbook, provide various checklists that can be applied to a diverse array of study designs, for different systematic review types [2, 3, 5–8].

Systematic reviews of diagnostic test accuracy (DTA) studies include evaluations of one or more index tests against a reference standard. Findings from such reviews are used by clinicians when deciding whether a medical test can identify patients with the target condition, or when facing a choice between two alternative tests. However, making a confident clinical decision based on a review of DTA studies can be challenging, since studies included in such reviews may suffer from methodological shortcomings, putting them at risk of bias [8, 9].

The current instrument for evaluating the methodological strength of DTA studies in systematic reviews is known as the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) tool. This tool covers four key domains: patient selection, index test, reference standard, and flow of patients through the study and timing of the index test(s) and reference standard [7, 8]. The authors' final judgments, based on this tool and other instruments, can be presented in reviews as either tables or figures. In Cochrane reviews, these can be created in Review Manager. The two figures that are found most often in systematic DTA reviews as a summary of the risk-of-bias assessment are as follows: a stacked bar chart, showing the proportion of studies with each of the judgments ("Low risk," "High risk," "Unclear risk" of bias) and a plot that presents all judgments as a cross-tabulation of studies against domains, usually called a "traffic light" plot [2, 7].

These figures can be presented not only for all studies included in the review but also per meta-analysis specifically. The advantage of presenting traffic light plots alongside forest plots for a specific meta-analysis is that the overall risk of bias for a specific summary estimate can be clear at a glance. Such a summary graph can be regarded as a visual representation of the credibility of the included evidence: the extent to which the included studies are believed to be at low risk of bias. This not only helps the reviewers to consider results of their risk-of-bias assessment when drawing conclusions, it can also help readers, by giving them a quick overview of the validity of the evidence within the review [7, 10]. With a fair and precise presentation of the validity of the studies included in a systematic review, readers will be able to appraise the certainty of the available evidence, a key element for evaluating whether the review findings support a particular clinical recommendation [11]. Cochrane encourages authors to use stratification by overall risk-of-bias judgment as the default strategy in meta-analyses of randomized trials but not for diagnostic test accuracy reviews. An example of a forest plot that displays domain specific risk-of-bias and overall risk-of-bias, with the meta-analysis stratified by overall risk-of-bias, can be seen in a figure presented by Sterne et al. [12].

Studies included in systematic reviews can vary substantially in total sample size and in the relative number of study participants with and without the target condition. These differences will affect summary estimates in meta-analysis, with larger studies typically contributing more to the summary estimates, and studies with more diseased patients having a larger effect on estimates of sensitivity [13–17]. This means that one should be more worried when one of the larger studies in a review is at high risk of bias, compared to a situation in which only a very small study is at high risk of bias. Yet, at present, summaries of risk-of-bias assessments are usually presented at the study level, with all studies contributing in a similar way to such summaries. Although some suggestions were made to use more informative methods of presenting risk-of-bias assessments, which could illustrate the relative contributions of studies with each of risk-of-bias judgment [2, 18], differences in absolute or relative sample size do not seem to be included in the current commonly used method, especially in diagnostic accuracy studies.

We here present alternative methods for summarizing risk-of-bias assessments in systematic reviews of diagnostic accuracy studies. The alternative methods draw more attention to the relative contribution of included studies to the review. By incorporating study sample size or effective sample size in the risk-of-bias summary, rather than just the number of studies, these alternative methods could provide a more informative depiction of the validity of the total body of evidence in the review.

Methods

Motivating example

We used existing systematic reviews of diagnostic accuracy studies as examples to illustrate the existing and

novel methods of the visual presentation of risk-of-bias. To demonstrate the generalizability of our findings, we selected four reviews that differ in the number of included studies (ranging from 9 to 32), across a variety of clinical domains.

Two systematic reviews targeted non-invasive tests in patients with non-alcoholic fatty liver disease (NAFLD). Studies were eligible if they included adult patients with biopsy-proven or suspected NAFLD for evaluating CK18 [19] or Enhanced Liver Fibrosis (ELF) test [20] as the index test, with liver biopsy as the reference standard. The target conditions were liver fibrosis and non-alcoholic steatohepatitis. One review included 32 reports of studies that had evaluated the diagnostic performance of CK18; the second review summarized 13 studies that had evaluated the ELF test.

The other two selected reviews are Cochrane systematic reviews, published in 2020. One systematic review targeted DTA studies evaluating the performance of measured hippocampal volume with structural magnetic resonance imaging for the early diagnosis of dementia due to Alzheimer's disease in people with mild cognitive impairment. Twenty-two studies were included in this systematic review [21]. The fourth systematic review aimed to assess the diagnostic accuracy of transcranial Doppler and transcranial color Doppler for detecting stenosis and occlusion of intracranial large arteries in people with acute ischemic stroke. This study included 9 DTA studies [22].

Reporting risk-of-bias assessment methods

The risk-of-bias assessment results of the four systematic reviews are presented in tables and illustrated in figures, using the current method and two alternative methods to show how the implementation of the new methods can alter the overall risk-of-bias assessment summary.

In all selected systematic reviews, two reviewers had used the QUADAS-2 tool to assess risk-of-bias and concerns about the applicability in the studies. In this report, we do not discuss possible consequences of our method for the concerns regarding applicability. We believe that applying these alternative methods to the risk-of-bias part of the four domains of QUADAS-2 tool could sufficiently illustrate the potential differences between the respective methods.

Current method

Using the commonly used risk-of-bias method, we generated bar graphs that display the proportion of studies with each of the risk-of-bias judgments for each of the four domains of the QUADAS-2 tool.

Weighted method—sample size

The commonly used risk-of-bias assessment and summary figures rely on the number of studies at the respective levels of risk-of-bias in each domain. This ignores the relative size of the included studies in the total risk-of-bias assessment. A study with a relatively large sample size contributes more to the review but is treated equally, compared to a study with a much smaller sample size.

The Cochrane handbook for systematic reviews of interventions recommends to present the risk-of-bias assessment results by restricting attention to studies in a particular importance to meta-analysis and to represent the proportion of information at different risk-of-bias levels [2]. However, such weighted plots are not producible in Cochrane's Review Manager.

It is very well possible to assign different weights to the studies when preparing summaries, to display how the included studies contribute to the total body of evidence in the review. One way to do so is using relative total sample size as the weight, which reveals the relative contribution of each study to the total group of patients for which data are included in the systematic review. Assigning differential weights to studies based on their relative sample size would be especially influential when considerable differences in sample size exist between included studies.

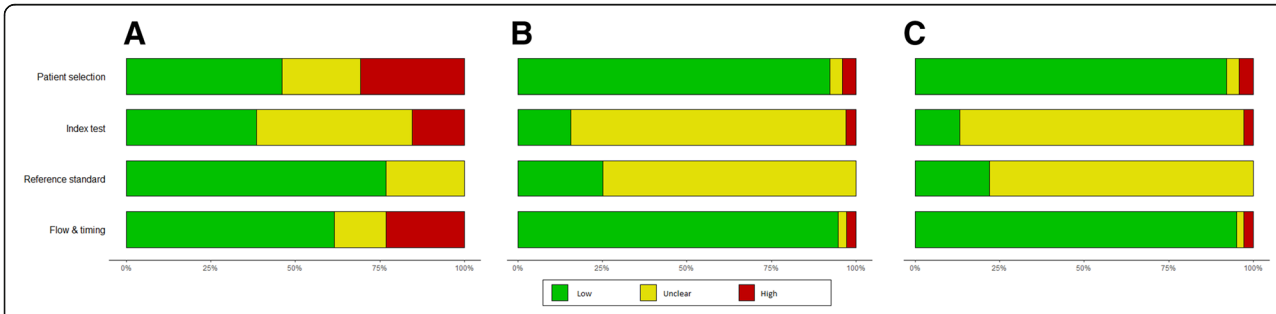

Fig. 1 Results of risk-of-bias assessment plots, which illustrate the judgments ("Low risk," "High risk," and "Unclear risk" of bias) for four QUADAS-2 tool domains (x-axis) based on (**A**) proportion of included studies, (**B**) proportion of included patients (**C**) effective sample size of 13 included studies in the ELF systematic review (y-axis)

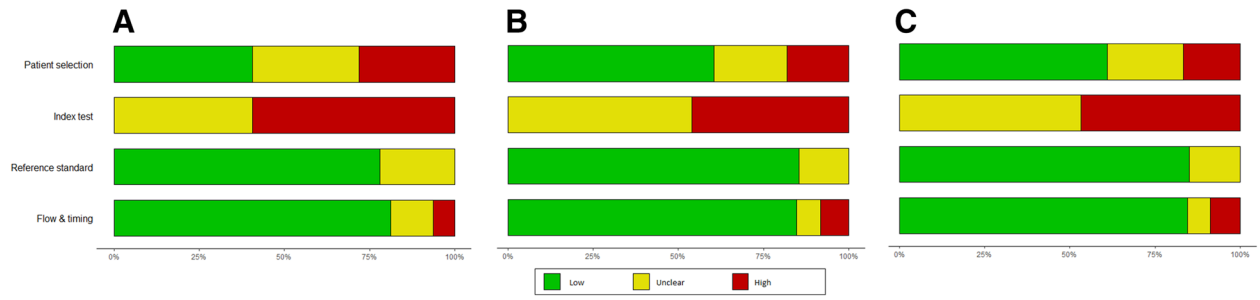

Fig. 2 Results of risk-of-bias assessment plots, which illustrate the judgments ("Low risk," "High risk," and "Unclear risk" of bias) for four QUADAS-2 tool domains (x-axis) based on (**A**) proportion of included studies, (**B**) proportion of included patients (**C**) effective sample size of 32 included studies in the CK18 systematic review (y-axis)

Accounting for differences in sample size in risk-of-bias assessment would bring this step of systematic reviews in line with methods for meta-analysis, which do not rely on vote counting on a study-by-study level, but incorporate the relative precision of each study in producing summary estimates. In general, recommended methods include inverse variance-weighted average methods or relying on weighted sums of z-scores [13]. Similar to these weighting methods for interventional studies, weighted average estimators are presented for meta-analysis of diagnostic test accuracy studies [23]. In DTA reviews, hierarchical methods, such as the bivariate logit-normal model, also account for between-study differences in sample size [24, 25].

Weighted method—effective sample size

Simple weighting by sample size may not be always sufficient [16, 17]. Study groups that are equal in size can include quite different numbers of participants with (n2) and without the target condition (n1). The proportion of cases with the target condition commonly differs across the various setting accuracy studies are conducted in. Consequently, these differences can affect the precision of an estimate of test accuracy for a given total sample size [16, 23].

An alternative is to rely on the effective sample size as a more appropriate method to display the relative contribution of a study. Deeks and his colleagues presented a simple formula for calculating effective sample size in DTA studies and stated in their report that "sample size related precision when there are unequal group sizes is more appropriately summarized by the effective sample size, where ESS= (4n1n2)/(n1 + n2)" [16].

After presenting the findings of the four systematic reviews based on the current risk-of-bias assessment method and the proportion of studies at low, unclear, and high risk of bias, we then used our new methods and replaced the proportion of studies with total sample size of individual studies and their effective sample size at different risk-of-bias levels [16]. Accordingly, we presented an alternative version of the graphs to present the summary, one that relies on the sample size and effective sample size of the included studies at different levels of risk-of-bias.

Results

The results of the current risk-of-bias assessment method are presented in Figs. 1, 2, 3, and 4A, which illustrate the proportion of studies at different risk levels. While the findings from the alternative weighting methods are illustrated in Figs. 1, 2, 3, and 4B and C. In the tables, we reported the findings as frequency and

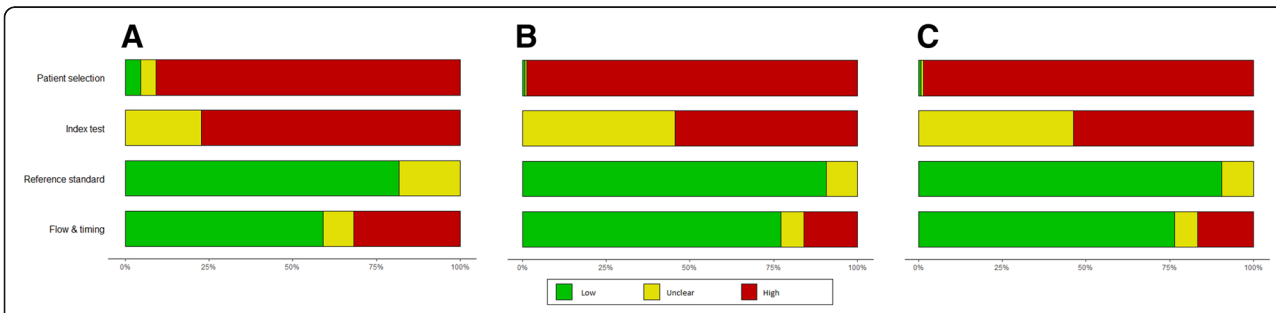

Fig. 3 Results of risk-of-bias assessment plots, which illustrate the judgments ("Low risk," "High risk," and "Unclear risk" of bias) for four QUADAS-2 tool domains (x-axis) based on (**A**) proportion of included studies, (**B**) proportion of included patients (**C**) effective sample size of 22 included studies in the Lombardi 2020 systematic review (y-axis)

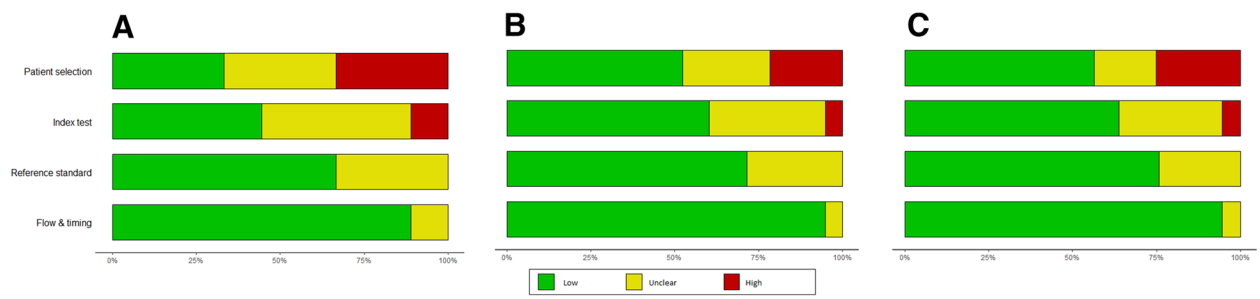

Fig. 4 Results of risk-of-bias assessment plots, which illustrate the judgments ("Low risk," "High risk," and "Unclear risk" of bias) for four QUADAS-2 tool domains (x-axis) based on (**A**) proportion of included studies, (**B**) proportion of included patients (**C**) effective sample size of 9 included studies in the Mattioni 2020 systematic review (y-axis)

percentage of low, unclear, and high risk of bias for each QUADAS-2 tool domain.

Figure 1A shows the summary risk-of-bias plot of studies that evaluated the performance of the ELF test in detecting liver fibrosis or NASH in NAFLD patients. This summary plot is based on the percentages of included studies. In contrast, Fig. 1B and C shows the assessment results when including study sample size or effective sample size, respectively. For the patient selection domain, the risk of bias was high in 31% of studies. However, after replacing the number of studies with the relative sample size and effective sample size of the individual studies, it changed to significantly smaller proportions (4%). The results in unclear and low-risk levels also changed when using alternative weighted methods: 23% vs 4% and 46% vs 92%.

In the other domains, a similar, considerable difference was observed between the results of non-weighted and weighted methods. For instance, in the index test domain, the percentage in the high-risk level changed from 15 to 3%. In the unclear and low-risk levels of this domain, differences were observed not only between the current risk-of-bias and the weighted methods but also between the two-weighted methods. The results changed from 46 to 38% in the first assessment to 81% and 16% using sample size, and from 84 to 13% when relying on effective sample size weighting method.

In the reference standard domain, there were no studies at high risk-of-bias. The 23% of studies for which risk-of-bias level was judged "unclear" changed to 75% of patients, after applying weights based on sample size. While at low risk-of-bias, the number changed from 77% of studies to 25% of population. The effective sample size weighting method resulted in 78% and 22% at unclear and low risk-of-bias, respectively.

The results in the flow and timing domain also changed from 23 to 3% in high-risk level, from 15 to 2% in unclear-risk level and from 62 to 95% in low-risk level after applying weights to the studies. See Table 1 for the details.

Using different weighting methods also showed noticeable changes in risk-of-bias assessment results for the other selected systematic reviews. See Figs. 2, 3, and 4 for the risk-of-bias summary plots before weighting (A) and after using weighted methods based on sample size (B) and effective sample size (C). Tables 2, 3, and 4 show the detailed changes in percentages of each level of bias in different QUADAS-2 domains. In general, the observed differences between the methods were more noticeable when the reviews included a smaller number of studies with wider range of sample size.

Discussion

We presented alternative methods to summarize the risk-of-bias assessments in systematic reviews of diagnostic test accuracy studies. By using these methods, including either relative sample size or relative effective sample size of the individual studies, we observed considerable visual changes for the four examples when presenting the risk-of-bias levels for each domain of the QUADAS-2 checklist, compared to the common

Table 1 Risk-of-bias (RoB) levels based on proportion of studies, their sample size, and effective sample size in ELF systematic review

QUADAS2 tool domain	RoB based on number of studies			RoB based on sample size			RoB based on effective sample size		
Risk of bias	Low	Unclear	High	Low	Unclear	High	Low	Unclear	High
Patient selection	46%	23%	31%	92%	4%	4%	92%	4%	4%
Index test	38%	46%	15%	16%	81%	3%	13%	84%	3%
Reference standard	77%	23%	0%	25%	75%	0%	22%	78%	0%
Flow and timing	62%	15%	23%	95%	2%	3%	95%	2%	3%

Table 2 Risk-of-bias (RoB) levels based on proportion of studies, their sample size, and effective sample size in CK18 systematic review

QUADAS2 tool domain	RoB based on number of studies			RoB based on sample size			RoB based on effective sample size		
Risk of bias	Low	Unclear	High	Low	Unclear	High	Low	Unclear	High
Patient selection	41%	31%	28%	60%	22%	18%	61%	22%	17%
Index test	0%	41%	59%	0%	54%	46%	0%	53%	47%
Reference standard	78%	82%	0%	85%	15%	0%	85%	15%	0%
Flow and timing	81%	13%	6%	85%	7%	8%	85%	7%	9%

unweighted method, which relies on the proportion of studies.

Systematic reviews and meta-analyses have become increasingly important in healthcare settings. Policy makers and clinicians rely on high quality systematic reviews for their decision-making. Yet, as a form of observational research, systematic reviews are susceptible to potential bias. When some of the included studies have methodological shortcomings, the meta-analytic results may be jeopardized [26, 27]. As studies included in a systematic review can be heterogeneous, also in terms of methodological rigor, they can, could, or should contribute in a different way to the total body of evidence, depending on their strengths and weaknesses [28].

Scores resulting from the risk-of-bias assessment could be used to weight the data of different studies included in a meta-analysis [29]. Work has been done in DTA systematic reviews on different methods of weighting studies according to their quality assessment result, to produce different risk-of-bias summaries, or to incorporate these in meta-analysis [30]. However, a common criticism of this approach is the lack of an empirical basis for deciding how much weight to assign to different domains of bias [2, 17, 31]. It has also been argued that calculating a summary score could lead to questionable assessments of validity [32] and that such scales may be less likely to present transparent summaries for review readers. For this reason, methodologists recommend avoiding direct weighting of effect estimates by risk-of-bias assessment results [2, 31].

We believe that meta-analysis is not the only phase in a systematic review that requires careful consideration of differences between included studies. Incorporating the methodological strength of the included studies in reports of reviews can and should influence conclusions drawn from the reviews. In a systematic review that included studies of different sizes and with methodological differences, studies that differ in their risk of bias should contribute differently to the total body of evidence. In our study, applying the alternative weighting methods illustrated how one large study at high risk of bias can be more influential in the total risk-of-bias assessment than a tiny study, also at high risk of bias. We believe methods for presenting risk-of-bias judgments that incorporate study weights can provide both authors and readers with more informative results of the risk-of-bias assessment. This will help in building valid conclusions and can facilitate decision-making based on the review findings.

Primary studies in a single systematic review may also have been performed in different settings and populations, with consequences for disease prevalence, even for studies with an identical sample size. Subsequently, differences in the relative balance of diseased and non-diseased study participants can affect precision of the accuracy estimates, for a given total sample size. Although we observed only small differences between total and effective sample size methods in our selected examples of systematic reviews, we believe that relying on effective sample size in summarizing risk-of-bias assessments, rather than on total sample size can be an even more informative weighting method, especially when the number of included primary studies is small and disease prevalence varies substantially [20, 22].

To facilitate the production of risk-of-bias assessment figures, a new Risk-Of-Bias VISualization tool, robvis, has recently been presented as an R package and a web app [18]. In this platform, a measure of the precision of the estimate, such as the weight assigned to that result in a meta-analysis or the study sample size, can be included to create the summary risk-of-bias plot. At

Table 3 Risk-of-bias (RoB) levels based on proportion of studies, their sample size, and effective sample size in Lombardi 2020

QUADAS2 tool domain	RoB based on number of studies			RoB based on sample size			RoB based on effective sample size		
Risk of bias	Low	Unclear	High	Low	Unclear	High	Low	Unclear	High
Patient selection	5%	5%	91%	1%	1%	99%	1%	1%	99%
Index test	0%	23%	77%	0%	46%	54%	0%	46%	54%
Reference standard	82%	18%	0%	91%	9%	0%	90%	10%	0%
Flow and timing	59%	9%	32%	77%	7%	16%	77%	7%	17%

Table 4 Risk-of-bias (RoB) levels based on proportion of studies, their sample size, and effective sample size in Mattioni 2020

QUADAS2 tool domain	RoB based on number of studies			RoB based on sample size			RoB based on effective sample size		
Risk of bias	**Low**	**Unclear**	**High**	**Low**	**Unclear**	**High**	**Low**	**Unclear**	**High**
Patient selection	33%	33%	33%	52%	26%	22%	56%	18%	25%
Index test	44%	44%	11%	60%	35%	5%	64%	31%	5%
Reference standard	67%	33%	0%	72%	28%	0%	76%	24%	0%
Flow and timing	89%	11%	0%	95%	5%	0%	95%	5%	0%

present, the package cannot yet produce graphs that show applicability concerns. Modifying bias domains within the tools is only possible for the "ROB1" option, which can handle varying numbers of columns, since authors using this tool frequently add or remove bias domains within this tool. Moreover, it is important to know how much awarding weights to the studies changes the risk-of-bias assessment findings, as in some levels the difference might be small and not recognizable in plots. We believe that the package could be further improved, providing percentages in the risk level at each domain, thereby helping authors in comparing weighted and unweighted methods and in interpreting the findings correctly.

Our examples were based on the QUADAS-2 risk-of-bias assessment tool for test accuracy studies. Future research could explore other risk-of-bias tools, as well as the impact on reviews with different levels of heterogeneity in included studies. It would also be informative to explore systematically to what extent systematic review authors and readers respond to these new weighted methods of risk-of-bias assessment.

Conclusion

We here have shown that an alternative way of summarizing risk-of-bias assessments with the QUADAS-2 tool can be used, one that does more justice to the relative contribution of each study to the total body of evidence included in the review. This can be achieved by using weights, either based on sample size or on effective sample size. We recommend reviewers select one of these alternative methods of weighting for summarizing the risk-of-bias assessment and to pre-specify the selected approach in the systematic review protocol, to avoid potential bias.

Evaluating and reporting the risk of bias in a review, thereby informing the readers about the limitations in the available body of evidence, will not be sufficient to produce valid conclusions. We call on reviewers to also incorporate the risk-of-bias assessment into their interpretation of the available data, their conclusions, and in the summary of findings. Only then we can trust that the conclusions in the review do justice to the validity of the research findings included in the systematic review.

Abbreviations

DTA: Diagnostic Test Accuracy; QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies 2; NAFLD: Non-Alcoholic Fatty Liver Disease; ELF: Enhanced Liver Fibrosis; RoB: Risk of Bias; ESS: Effective Sample Size; NASH: Non-Alcoholic Steatohepatitis

Acknowledgements

The authors sincerely thank Dr. Nahid Mostafavi for kindly helping us with illustrating the findings.

Authors' contributions

YV and PB contributed in designing the study. YV prepared the application of the method and the results. YV and PB prepared the first draft of the manuscript. Interpretation has been performed by PB and YV and ML. YV, PB, and ML reviewed and critically revised the manuscript. The authors read and approved the final manuscript.

Funding

This work has been supported by the LITMUS (Liver Investigation: Testing Marker Utility in Steatohepatitis) project, funded by the Innovative Medicines Initiative (IMI2) Program of the European Union (Grant Agreement 777377).

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Received: 8 December 2020 Accepted: 12 June 2021
Published online: 27 June 2021

References

1. Burns PB, Rohrich RJ, Chung KC. The levels of evidence and their role in evidence-based medicine. Plast Reconstr Surg. 2011;128(1):305–10. https://doi.org/10.1097/PRS.0b013e318219c171.
2. Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane handbook for systematic reviews of interventions: John Wiley & Sons; 2019. https://doi.org/10.1002/9781119536604.
3. Pussegoda K, Turner L, Garritty C, Mayhew A, Skidmore B, Stevens A, et al. Systematic review adherence to methodological or reporting quality. Syst Rev. 2017;6(1):1–14.
4. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol. 2003;3(1):25. https://doi.org/10.1186/1471-2288-3-25.
5. Clarke M. The Cochrane Collaboration and systematic reviews. Br J Surg. 2007;94(4):391–2. https://doi.org/10.1002/bjs.5812.
6. Viswanathan M, Patnode CD, Berkman ND, Bass EB, Chang S, Hartling L, et al. Assessing the risk of bias in systematic reviews of health care interventions.

Methods guide for effectiveness and comparative effectiveness reviews: Agency for Healthcare Research and Quality (US); 2017.
7. Macaskill P, Gatsonis C, Deeks J, Harbord R, Takwoingi Y. Cochrane handbook for systematic reviews of diagnostic test accuracy. Version 09 0. London: The Cochrane Collaboration; 2010.
8. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529–36. https://doi.org/10.7326/0003-4819-155-8-201110180-00009.
9. Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM. Systematic reviews of diagnostic test accuracy. Ann Intern Med. 2008;149(12):889–97. https://doi.org/10.7326/0003-4819-149-12-200812160-00008.
10. Ochodo EA, Van Enst WA, Naaktgeboren CA, De Groot JA, Hooft L, Moons KG, et al. Incorporating quality assessments of primary studies in the conclusions of diagnostic accuracy reviews: a cross-sectional study. BMC Med Res Methodol. 2014;14(1):33. https://doi.org/10.1186/1471-2288-14-33.
11. Hultcrantz M, Rind D, Akl EA, Treweek S, Mustafa RA, Iorio A, et al. The GRADE Working 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.
12. Sterne J, Savović J, Page M, Elbers R, Blencowe N, Boutron I, et al. RoB 2: A revised Cochrane risk-of-bias tool for randomized trials. BMJ. 2019;366: l48981.
13. Lee CH, Cook S, Lee JS, Han B. Comparison of two meta-analysis methods: inverse-variance-weighted average and weighted sum of Z-scores. Genom Inform. 2016;14(4):173–80. https://doi.org/10.5808/GI.2016.14.4.173.
14. Marín-Martínez F, Sánchez-Meca J. Weighting by inverse variance or by sample size in random-effects meta-analysis. Educ Psychol Meas. 2010;70(1): 56–73. https://doi.org/10.1177/0013164409344534.
15. Sánchez-Meca J, Marin-Martinez F. Weighting by inverse variance or by sample size in meta-analysis: a simulation study. Educ Psychol Meas. 1998; 58(2):211–20. https://doi.org/10.1177/0013164498058002005.
16. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. J Clin Epidemiol. 2005;58(9):882–93. https://doi.org/10.1016/j.jclinepi.2005.01.016.
17. Gatsonis C, Paliwal P. Meta-analysis of diagnostic and screening test accuracy evaluations: methodologic primer. Am J Roentgenol. 2006;187(2): 271–81. https://doi.org/10.2214/AJR.06.0226.
18. McGuinness LA, Higgins JP. Risk-of-bias VISualization (robvis): an R package and Shiny web app for visualizing risk-of-bias assessments. Res Synth Methods. 2020.
19. Lee J, Vali Y, Boursier J, Duffin K, Verheij J, Brosnan MJ, et al. Accuracy of cytokeratin 18 (M30 and M65) in detecting non-alcoholic steatohepatitis and fibrosis: a systematic review and meta-analysis. PLoS ONE. 2020:1–19.
20. Vali Y, Lee J, Boursier J, Spijker R, Löffler J, Verheij J, et al. Enhanced liver fibrosis test for the non-invasive diagnosis of fibrosis in patients with NAFLD: a systematic review and meta-analysis. J Hepatol. 2020;73(2):252–62. https://doi.org/10.1016/j.jhep.2020.03.036.
21. Lombardi G, Crescioli G, Cavedo E, Lucenteforte E, Casazza G, Bellatorre AG, et al. Structural magnetic resonance imaging for the early diagnosis of dementia due to Alzheimer's disease in people with mild cognitive impairment. Cochrane Database Syst Rev. 2020;3.
22. Mattioni A, Cenciarelli S, Eusebi P, Brazzelli M, Mazzoli T, Del Sette M, et al. Transcranial Doppler sonography for detecting stenosis or occlusion of intracranial arteries in people with acute ischaemic stroke. Cochrane Database Syst Rev. 2020;2.
23. McClish DK. Combining and comparing area estimates across studies or strata. Med Decis Mak. 1992;12(4):274–9. https://doi.org/10.1177/0272989X9201200405.
24. Harbord RM, Whiting P, Sterne JA, Egger M, Deeks JJ, Shang A, et al. An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. J Clin Epidemiol. 2008;61(11): 1095–103. https://doi.org/10.1016/j.jclinepi.2007.09.013.
25. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med. 2001;20(19):2865–84. https://doi.org/10.1002/sim.942.
26. Whiting P, Savović J, Higgins JP, Caldwell DM, Reeves BC, Shea B, et al. ROBIS: a new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol. 2016;69:225–34. https://doi.org/10.1016/j.jclinepi.2015.06.005.
27. Leeflang M, Reitsma J, Scholten R, Rutjes A, Di Nisio M, Deeks J, et al. Impact of adjustment for quality on results of metaanalyses of diagnostic accuracy. Clin Chem. 2007;53(2):164–72. https://doi.org/10.1373/clinchem.2006.076398.
28. Burke DL, Ensor J, Snell KI, van der Windt D, Riley RD. Guidance for deriving and presenting percentage study weights in meta-analysis of test accuracy studies. Res Synth Methods. 2018;9(2):163–78. https://doi.org/10.1002/jrsm.1283.
29. La Torre G, Chiaradia G, Gianfagna F, Boccia S, De Laurentis A, Ricciardi W. Quality assessment in meta-analyses. Ital J Public Health. 2006;3(2).
30. Whiting P, Harbord R, Kleijnen J. Scoring the quality of diagnostic accuracy studies: an example using QUADAS; 2004.
31. Greenland S, O'rourke K. On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. Biostatistics. 2001; 2(4):463–71. https://doi.org/10.1093/biostatistics/2.4.463.
32. Jüni P, Altman DG, Egger M. Assessing the quality of controlled clinical trials. Bmj. 2001;323(7303):42–6. https://doi.org/10.1136/bmj.323.7303.42.

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

BMC