Journal of Clinical Epidemiology 143 (2022) 137–148

REVIEW

Quality assessment practice in systematic reviews of mediation studies: results from an overview of systematic reviews

Tat-Thang Vo[a,*], Aidan Cashin[b,1], Cecilia Superchi[c,1], Pham Hien Trang Tu[d,1], Thanh Binh Nguyen[e,1], Isabelle Boutron[c], David MacKinnon[f], Tyler Vanderweele[g], Hopin Lee[h,2], Stijn Vansteelandt[i,2]

[a] Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia
[b] Centre for Pain IMPACT, Neuroscience Research Australia, Sydney, Australia
[c] CRESS, INSERM, INRA, Université de Paris, Paris, France
[d] Department of Family Medicine and Population Health, University of Antwerp, Antwerp, Belgium
[e] Department of Pharmacy, Vietnam National Cancer Hospital, Hanoi, Vietnam
[f] Department of Psychology, Arizona State University, Tempe
[g] Departments of Epidemiology and Biostatistics, Harvard School of Public Health, Harvard University, Boston
[h] Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, United Kingdom
[i] Department of Applied Mathematics, Computer Science and Statistics, Faculty of Science, Ghent University, Ghent, Belgium

Accepted 7 December 2021; Available online 13 December 2021

Abstract

Objective: To describe the bias assessment practice in recently published systematic reviews of mediation studies and to evaluate the quality of different bias assessment tools for mediation analysis proposed in the literature.

Method: We conducted an overview of systematic reviews by searching MEDLINE (OvidSP), PsycINFO (OvidSP), Cochrane Database of Systematic Reviews (OvidSP), and PubMed databases for systematic reviews of mediation studies published from 2007 to 2020. Two reviewers independently screened the title, abstracts, and full texts of the identified reports and extracted the data. The publications of all mediation-specific quality assessment tools used in these reviews were also identified for the evaluation of the tools' development and validation.

Result: Among 103 eligible reviews, 24 (23%) reviews did not assess the risk of bias of eligible studies, and 48 (47%) assessed risk of bias using a tool that was not specifically designed to evaluate mediation analysis. 31 (30.1%) reviews assessed the risk of mediation-specific biases, either narratively or by using specific tools for mediation studies. However, none of these tools were consensus-based, rigorously developed or validated.

Conclusion: The quality assessment practice in recently published systematic reviews of mediation studies is suboptimal. To improve the quality and consistency of risk of bias assessments for mediation studies, a consensus-based bias assessment tool is needed.

Conflict of interests: None declared.

[1] These authors equally contributed as second co-authors.

[2] These authors equally contributed as last co-authors.

[*] Corresponding author.: Tel.: (215) 898-8222; Fax: (215) 898-1280.
E-mail address: tatthang@wharton.upenn.edu (T.-T. Vo).

https://doi.org/10.1016/j.jclinepi.2021.12.013
0895-4356/© 2021 Elsevier Inc. All rights reserved.

What is new?

Key findings

- In addition to standard sources of biases in RCTs and observational studies, mediation analyses are prone to other sources of bias, such as temporal order bias and mediator-outcome confounding bias.
- In practice, only about 30% of mediation systematic reviews assessed the risk of mediation-specific biases, either narratively or by using specific tools for mediation studies.
- None of the mediation-specific quality assessment tools identified in the literature were consensus-based, rigorously developed or validated.

What this adds to what is known

- The quality assessment practice in recently published systematic reviews of mediation studies is suboptimal, which increases the risk of mediation-specific biases not being properly evaluated.

What is the implication and what should change now

- To improve the quality and consistency of risk of bias assessments for mediation studies, a consensus-based bias assessment tool for mediation analysis is needed.

1. Introduction

Mediation analysis is a very common type of statistical analysis in psychology, sociology, epidemiology, and medicine [1,2]. Mediation analyses of randomized controlled trials (RCTs) and observational studies in health and medical research can generate evidence about the relative magnitude of different pathways and mechanisms by which an exposure may affect an outcome [2,3]. Through mediation analysis, the total effect of an exposure on an outcome can be decomposed into an indirect effect that works through a mechanism(s) of interest, and a direct effect that works through any other mechanisms. Systematic reviews and meta-analyses of mediation studies are increasingly being implemented in health and medical research [4,5]. These reviews aim to summarize all available evidence on the role of one or several mediators in explaining a specific treatment/exposure – outcome relationship [4,5].

As in systematic reviews of (RCTs) and observational studies, an important step in systematic reviews of mediation analyses is to assess the potential risk of bias in each eligible study. In addition to the sources of biases common to RCTs and observational studies, mediation analyses are prone to additional sources of bias, for instance, temporal order bias which occurs when mediators are not measured prior to the outcome and after treatment completion, or mediator-outcome confounding bias which may occur even in high quality RCTs [2,4].

It is unclear how recently published systematic reviews of mediation studies conduct quality assessments. The most recent evidence from a sample of systematic reviews of mediation studies published up until March 2017 highlighted the inconsistency in quality assessment conduct and heterogeneity in the choice of quality assessment tool [5]. Considering continued methodological advances in mediation methods, and the rapid increases in mediation publications [6], an update is needed.

The most recent quality assessment tools for RCTs (e.g. the ROB 2.0 tool [7]) or observational studies (e.g. the ROBIN-I tool [8]) which assess bias across a range of domains (e.g., bias due to deviations from intended interventions and bias due to missing data) have not been adapted or extended to assess mediation-specific biases. Without consolidated guidance for assessing the risk of bias in mediation studies, it is unclear how mediation-specific biases are being assessed in systematic reviews and whether they are methodologically valid.

In this review, we aim to update the previous review by Cashin et al 2020 [5] to provide a snapshot of the quality assessment practice in recently published systematic reviews of mediation studies. A second aim is to assess the methodological quality of the mediation-specific bias assessment tools used in these reviews and to describe the investigated biases.

2. Methods

2.1. Study design

We conducted a methodological systematic review updating and extending the previous review by Cashin et al 2020 [5]. This review was reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [9]. The protocol was not registered in The International Prospective Register of Systematic Reviews because this review does not contain direct health-related outcomes [10].

2.2. Eligibility criteria

We adapted the eligibility criteria used by Cashin et al (2020) in a recent overview of mediation systematic reviews [5]. We included articles that considered systematic methods to identify primary studies which conducted a formal mediation analysis (e.g., product or difference in coefficients, latent growth modelling, causal mediation analysis) to investigate the mechanisms of health interventions or exposures on human participants of any age. Systematic reviews that (i) did not include studies that report an indirect effect, or (ii) only reported an exposure-mediator effect or mediator-outcome effect but not both were excluded. We also excluded (iii) non-English publications and (iv) protocols of systematic reviews. The full eligibility criteria are available in Appendix 1.

2.3. Information sources and search strategy

We used the search strategy developed by Cashin et al (2020) [5]. Previously, these authors searched MEDLINE (OvidSP), PsycINFO (OvidSP), Cochrane Database of Systematic Reviews (OvidSP), and PubMed databases for eligible systematic reviews published from 2007 to 2017. We made use of their results and updated the search for eligible systematic reviews published from 2017 to July 13, 2020 on the same databases. We hand-searched the reference lists of included studies to further identify other eligible articles that were not detected through the database search. The full search strategy is available in Appendix 2.

2.4. Study selection

We downloaded the search results into EndNote^{TM} and exported them to Microsoft Office Excel. Two reviewers (T.B.N. and H.T.T.P.) independently screened the titles and abstracts of all retrieved references to identify the eligible reports. The full-text copies of potentially eligible reports were also obtained and independently examined for further assessment if needed. In the case of disagreement, a third reviewer (T.T.V) was asked for opinion. The result of this process was reported through a PRISMA flowchart [9].

In what follows, we identified all mediation-specific quality assessment tools used in the selected systematic reviews. We defined a mediation-specific quality assessment tool as any instrument used to assist reviewers to assess and summarize the methodological quality of the eligible mediation studies. Once the eligible systematic reviews were identified, two reviewers (T.T.V and A.C) selected all mediation-specific quality assessment tools used in these reviews for evaluation. The same reviewers then conducted a specific search for the original publications describing the development of the identified tools. Three other reviewers (H.T.T.P, T.B.N and C.S) independently double-checked 50% of the process of selecting the mediation-specific tools from the eligible studies. In case of disagreement, consensus was determined by a discussion between the reviewers.

2.5. Data extraction

2.5.1. Quality assessment practice

A data-extraction form was designed, pilot-tested, and refined by a reviewer (T.T.V.) to extract the following information from the eligible reports: (1) characteristics of the systematic reviews, (2) whether the systematic review assessed the quality of each eligible study and if so, which quality assessment approach was used.

For (1), we determined the study design, health care field, intervention/exposure type, number and type of mediators, outcome type and data synthesis method. For (2), we classified the studies into three categories. A: use of a general quality assessment tool not specific to mediation; B: use of a mediation-specific quality assessment tool; and C: no use of a quality assessment tool but a narrative assessment of biases. Reviews could be classified into multiple categories.

2.5.2. Quality/Risk of bias assessment tool evaluation

A second data-extraction form was designed to extract the general characteristics of the mediation-specific quality assessment tools from their publications. The form was developed by a reviewer (T.T.V.) after consulting Superchi et al (2019) [11]. We first determined whether the tool was a scale or checklist (scale operationalized as a tool that produces a numeric or nominal overall quality score, and as a checklist otherwise). We recorded (i) the total number of items, (ii) how items were weighted (if any), (iii) how the overall score was calculated, (iv) the scoring range, (v) whether the scoring instructions were defined, (vi) the development, validation, assessment of the tool's reliability and (vii) the scope of the tool (i.e., whether the tool assessed internal, external validity or reporting quality). Full data extraction was conducted by two reviewers (T.T.V and A.C). Three other reviewers independently double-checked half of the extracted articles (H.T.T.P, T.B.N, C.S). In case of disagreement, consensus was determined by a discussion between the reviewers in charge of extracting the article with disagreement. Authors of the reports were contacted up to two times over a period of four weeks in cases where we needed further clarification. If contact was not possible, we excluded the tool from the analysis.

2.6. Data synthesis

Two reviewers (T.T.V and A.C) independently classified all items of each mediation-specific quality assessment tool into discrete quality domains (i.e, study design, non-mediation-specific bias, mediation-specific bias and non-bias-related aspect). In case of disagreement, we invoked input from a third reviewer (H.L). Once the list of quality domains was agreed upon, we determined the proportional contribution of different domains in each quality assessment tool, based on the number of items in the tool that belonged to each domain. With the proportions obtained, we created a domain profile for each tool. Then, we calculated the matrix of Euclidean distances between the domain profiles. These distances were used to perform the hierarchical, complete-linkage clustering analysis, which provided us with a tree structure that identified the domain similarities among the tools.

Categorical data were summarized using frequencies and percentages. Continuous data were summarized using

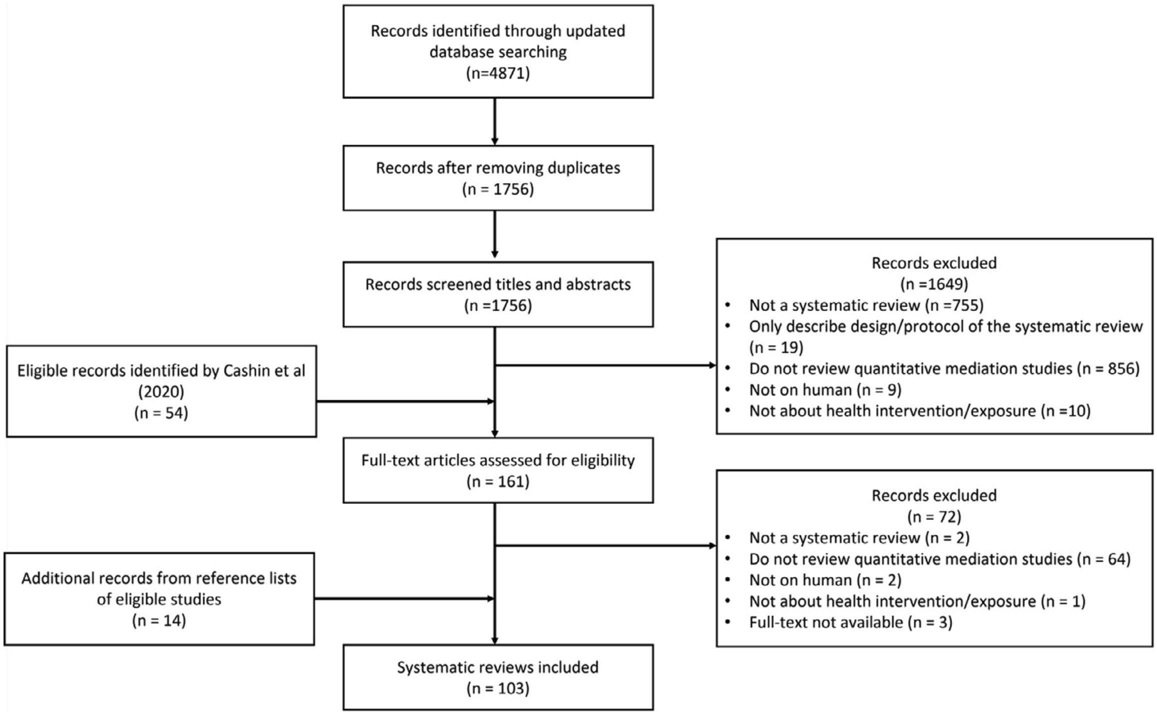
Records identified through updated database searching (n=4871)
Records after removing duplicates (n = 1756)
Records screened titles and abstracts (n =1756)
Records excluded (n =1649)
• Not a systematic review (n =755)
• Only describe design/protocol of the systematic review (n = 19)
• Do not review quantitative mediation studies (n = 856)
• Not on human (n = 9)
• Not about health intervention/exposure (n =10)
Eligible records identified by Cashin et al (2020) (n = 54)
Full-text articles assessed for eligibility (n = 161)
Records excluded (n = 72)
• Not a systematic review (n = 2)
• Do not review quantitative mediation studies (n = 64)
• Not on human (n = 2)
• Not about health intervention/exposure (n = 1)
• Full-text not available (n = 3)
Additional records from reference lists of eligible studies (n = 14)
Systematic reviews included (n = 103)

Fig. 1. Study selection PRISMA Flowchart.
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

median and interquartile range. Data were analyzed using MS Excel 2010 and R version 3.3.3.

3. Results

3.1. Study selection

The PRISMA flow diagram summarizing the screening process is presented in Fig. 1. Of 1824 references identified, 1721 reports were excluded for the following reasons: not a systematic review (n = 757), protocol of a systematic review (n = 19), eligible studies did not consider a formal mediation analysis (n = 920), non-human participants (n = 11), full-text unavailable (n = 13) and did not assess a heath intervention or exposure (n = 10).

3.2. Current quality assessment practice in systematic reviews of mediation studies

Characteristics of 103 eligible reviews are summarized in Table 1. The quality assessment in these reviews is described in Fig. 2 and appendix 2. Among these 103 eligible reviews, 76.7% (n = 79) assessed the methodological quality of the eligible mediation studies, mostly by a single quality assessment approach (68.9%, n = 71). 46.6% (n = 48) of reviews used a quality assessment tool that did not evaluate mediation-specific biases (e.g., the Newcastle-Ottawa Scale [12], the Scottish Intercollegiate Guidelines Network Methodology checklist for RCTs [13], The Cochrane Risk of Bias Assessment Tool for Non-Randomized Studies of Interventions [14] and so forth). In contrast, a mediation-specific quality assessment tool was only used in 11.7% (n = 12) of reviews. A few other reviews (10.7%, n = 11) did not assess the quality of each eligible study separately, but instead described the prevalence across studies of multiple methodological characteristics that affected the validity of the mediation findings. The list of these methodological characteristics is provided in appendix 3.

3.3. Characteristics of the mediation-specific quality assessment tools

From the 103 systematic reviews, we identified 12 tools, including ten scales, one checklist and one domain-based tool. The tools were developed from 2008 to 2019. Table 2 presents the general characteristics of the identified tools. None of the tools defined its scope or described its validity and reliability. None of the tools were consensus-based. Eight tools were developed by adapting a previously proposed tool. Most tools were used in only one systematic review identified in the previous step.

Nine tools (eight scales and one checklist) used binary questions with each score receiving an equal weight of one. The number of binary questions ranged from 7 to 16. The total score represented the sum of all responses coded "yes", and the proportion was the sum divided by the total number of questions. In six of eight scales with a sum

Table 1. Characteristics of 103 eligible systematic reviews

Characteristics	**N**	**%**
Field		
Mental health	39	37.9%
Behavioural medicine	26	25.2%
Bariatrics	10	9.7%
Musculoskeletal	6	5.8%
Cardio-metabolic health	4	3.9%
Addiction	4	3.9%
Others	14	13.6%
Design, n (%)		
Only included randomized controlled trials	34	33.0%
Only included observational studies	18	17.5%
Combination	51	49.5%
Data synthesis method, n (%)		
Narrative only	72	69.9%
Quantitative only	2	1.9%
Combination	29	28.2%
Number of mediators investigated, n (%)		
Single mediator	16	15.5%
Multiple mediators	87	84.5%
Exposure/Intervention category		
Exposure (in reviews of observational studies)		
Psychological, cognitive or behavioral exposure	29	28.2%
Sociology (e.g., socioeconomic status, stigma, etc.)	20	19.4%
Onset of a medical condition (e.g., pain, obesity, ADHD, etc.)	11	10.7%
Others	6	5.8%
Intervention (in reviews of randomized controlled trials)		
Psychological, cognitive or behavioral intervention	32	31.1%
Communal support (e.g., community health worker-delivered intervention)	2	1.9%
Medication	1	1.0%
Multiple	1	1.0%
Other	1	1.0%
First primary outcome category		
Behaviors (e.g., physical activity, alcohol use, dietary, etc.)	34	33.0%
Symptoms of mental disorders	28	27.2%
Physical & physiological functionality	10	9.7%
Multiple (e.g., asthma management, drinking outcomes, child development outcomes)	8	7.8%
Onset of obesity	5	4.9%
Psychological functioning and wellbeing	4	3.9%
Other	14	13.6%

or proportion of scores calculated, an overall conclusion about the quality of the assessed study (e.g. low, moderate or high quality) was established based on the obtained sum or proportion of scores (Table 2). However, none of the tools provided any instruction about how to interpret such conclusion (e.g. what is meant by an overall conclusion of *low quality*).

Two scales used multiple-choice questions. The options for each question were assigned a score. In one scale, the sum and proportion of scores were used to summarize these responses and to derive the conclusion about the quality of the assessed study. In the other scale, the overall assessment was based on the combination of the responses to the two questions included in the tool (e.g. a study was of low quality if it scored 2 in the first question and 3 in the second question, etc.). As above, none of the tools provided guidance on how to interpret the overall conclusion (Table 2).

Table 2. Characteristics of mediation quality assessment tools in the literature

Journal	First Author, Year	Format	Scope defined	OA[*]	Item & weights[†]	Scoring range[‡]	SSI[§]	Dev/Val/Rel[‖]	Freq[¶]	Details on the tool construction
Preventive Medicine	Lubans, 2008[16]	Scale	No	SS: 0-3: L; 4-6: M; 7-8: H	8, BQ Equal	0 – 8	ND	NA	1	Proposed by the authors
Journal of Nutrition Education and Behavior	Cerin, 2009[17]	Scale	No	SS: 0-3: L; 4-6: M; 7-9: H	9, BQ Equal	0 – 9	ND	NA	1	Modified from the tool of Lubans et al (2008)[16]
International Journal of Behavioral Nutrition and Physical Activity	Rhodes, 2010[18]	Scale	No	SS: 0-4: L; 5-8: M; 9-11: H	11, BQ Equal	0 – 11	ND	NA	4	Adding into the tool by Lubans et al[16] three additional items proposed by Cerin et al[17].
International Journal of Obesity	van Stralen, 2011[19]	Scale	No	PS: 0-70%: L; 70-100%: H	10, BQ Equal	0 – 100%	ND	NA	1	Combining the checklist of Lubans et al[16] and Cerin et al[17] and criteria proposed in a Delphi-based criteria list for QA of RCTs
Best Practice & Research: Clinical Rheumatology	Mansell, 2013[20]	Check-list	No	-	12, BQ -	-	-	NA	1	Adapting the tool of Lubans et al[16], Cerin et al[17] and Rhodes et al[18] by adding some new items recommended in the literature, following a methodological (non-systematic) review that they conducted.
Clinical Psychology Review	Gu, 2015[21]	Scale	No	SS: 0-5: L; 6-11: M; 12-16: H	16, BQ Equal	0 – 16	ND	NA	1	Adapting the tool of Lubans et al[16] by the CONSORT checklist, the Jadad checklist and Kazdin (2007)'s design requirements for mediation
Pain	Lee, 2015 [22]	Scale	No	SS: No rule for conclusion	7, BQ Equal	0 – 7	ND	NA	2	Modified from the tool of Mansell et al[20]
British Journal of Health Psychology	Windgassen, 2017[23]	Scale	No	SS/PS No rule for conclusion	8, BQ Equal	0 – 8/ 0 – 100%	ND	NA	1	Adding into the tool of Lubans et al[16] some items based on the standards of MA proposed by MacKinnon (2008)
Clinical Psychology Review	Hoppen, 2018[24]	Scale	No	SS/PS No rule for conclusion	12, MCQ Unequal	0 – 40/ 0 – 100%	ND	NA	1	Proposed by the authors
Clinical Psychology Review	Williams, 2018[15]	Domain-based	No	NA	7, DM -	NA	ND	NA	1	Adapted from the Effective Public Health Practice Project tool (EPHPP; Thomas, 2003)
Appetite	Claassen, 2019[25]	Scale	No	SS: D3-4/M3-5: B D2-M3-5 or D3-4/M2: M Otherwise: L	2, MCQ Equal	2 – 9	ND	NA	1	Proposed by the authors

(*continued on next page*)

Table 2 (*continued*)

Journal	First Author, Year	Format	Scope defined	OA*	Item & weights†	Scoring range‡	SSI§	Dev/Val/Rel‖	Freq¶	Details on the tool construction
PloS One	Cortés Garcías, 2019 [26]	Scale	No	SS: 0-4: W; 5-7: M; 8-9: S	9, BQ Equal	0 – 9	ND	NA	1	Adapted from the tool by Lee et al [22] with four items added in view of ‘standard guidelines’ for QA of RCTs and observational studies

* overall quality assessment (OA) by sum of score (SS) or proportion of score (PS). Based on the scores, studies are classified as having low/moderate/high quality (L-M-H ranking); or low/high quality (L-H ranking); or weak/moderate/strong evidence (W-M-S ranking). In the tool by Claassen et al (2019), there are 2 questions assessing 2 criteria, i.e. study design (D, score from 1 to 4) and mediation approach (M, score ranging from 1 to 5). Studies are classified by this tool to different levels of overall strength of mediation evidence, based on the response to the 2 questions: ’best’ – B for (D3/D4 and M3/M4/M5); ’moderate’ – M for (D3/D4 and M2) or (D2 and M3/M4/M5); ’low’ – L otherwise. (-): not applicable (i.e. when the tool is a checklist)

† number of items + whether items are simple binary questions (BQ), multiple-choice questions (CQ) or bias domains (DM) + if a score is calculated, whether items are equally weighted (Equal) or not (Unequal). (-): not applicable

‡ the minimum and maximum of the score for one study assessed by a scale. (-): not applicable

§ Definition of Scoring System Instruction (SSI) (e.g. what is meant by having low, moderate or high quality?). ND: not defined. (-): not applicable

‖ Information on the development (Dev), Validation (Val) and Reliability (Rel) of the tool. NA: information not available

¶ Frequency (Freq) of the tool being used among the identified systematic reviews of mediation studies. NA: information not available

In the domain-based tool, the signaling questions were not accessible even after contacting the authors. Each domain was rated by ’strong’, ’moderate’ and ’weak’. No overall bias rating was suggested by the authors.

3.4. Bias domains assessed in the tools and domain similarities across tools

The quality domains investigated in eleven tools with accessible content are provided in Table 3, Table 4 and

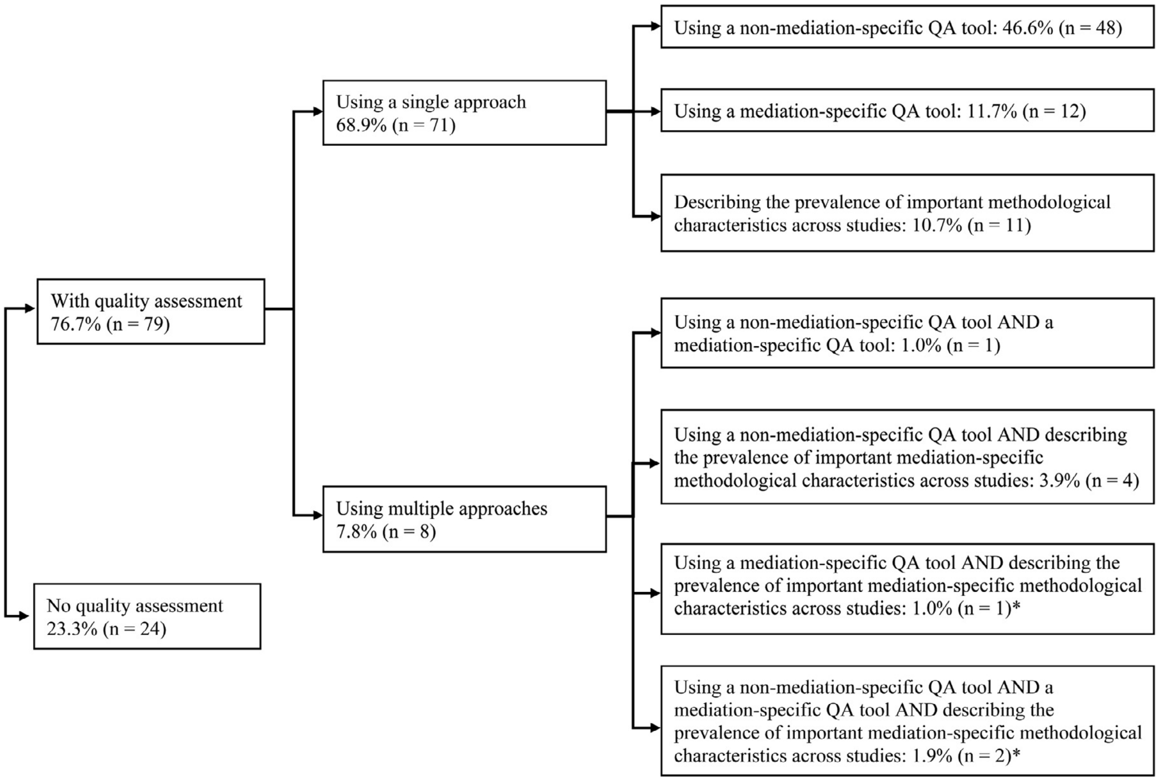
Using a non-mediation-specific QA tool: 46.6% (n = 48)
Using a single approach 68.9% (n = 71)
Using a mediation-specific QA tool: 11.7% (n = 12)
Describing the prevalence of important methodological characteristics across studies: 10.7% (n = 11)
With quality assessment 76.7% (n = 79)
Using a non-mediation-specific QA tool AND a mediation-specific QA tool: 1.0% (n = 1)
Using a non-mediation-specific QA tool AND describing the prevalence of important mediation-specific methodological characteristics across studies: 3.9% (n = 4)
Using multiple approaches 7.8% (n = 8)
Using a mediation-specific QA tool AND describing the prevalence of important mediation-specific methodological characteristics across studies: 1.0% (n = 1)*
No quality assessment 23.3% (n = 24)
Using a non-mediation-specific QA tool AND a mediation-specific QA tool AND describing the prevalence of important mediation-specific methodological characteristics across studies: 1.9% (n = 2)*

Fig. 2. Current quality assessment practice in systematic reviews of mediation studies. (*) the mediation-specific characteristics described were not discussed in the mediation-specific QA tool. QA: quality assessment

Table 3. Examples of signalling questions across 11 different quality/risk of bias assessment tools for mediation analysis

No.	Domains (n,%)	Examples
1. Study design (n = 10, 90.9%)		
1.1	Study design (n = 10, 90.9%)	- Did the study have an active control group?[21] - Study design (i.e. cross-sectional design → score 1; matched-controls design, retrospective cohort, case-control design, RCT or single-blind, non-randomized, placebo-controlled, repeated-measures design → score 3; Prospective cohort → score 5)[24]
2. Methodological bias (n = 11, 100%)		
2.1	Non-mediation-specific bias	
2.1.1	Randomization; Performance bias; Measurement bias; Bias due to missing data (n = 7, 63.6%)	- Was the method used to generate the sequence of randomization described and appropriate (table of random numbers, computer-generated, etc)?[21]; Were participants or experimenters blind to treatment assignment?[21] - Did the analysis include an intention-to-treat analysis?[19] - Was mediation analysis carried out using only the participants who received an adequate dose of the treatment?[21] - Were the psychometric characteristics of the mediator and outcome variables reported and were they within accepted ranges?[16,17,19] (e.g. reliability > 0.7[19], Cronbach's alpha/test-retest reliability > .7[21] or >.6[20]? - Was the drop-out rate non-selective?[19]
2.2	Mediation-specific bias	
2.2.1	Temporal order bias (n = 8, 72.7%)	- Did the study ascertain whether changes in the exposure variable preceded changes in the mediator variable?[20,22]; and whether changes in the mediating variables preceded changes in the outcome variables?[18,22]
2.2.2	Confounding bias (n = 7, 63.6%)	- Was post-intervention outcome controlled for baseline outcome?[21] - Did the study control for possible confounding factors? (Variables that may impact on results are identified and controlled for in terms of statistical analysis)[26]
2.2.3	Bias due to the inappropriate use of a statistical approach to investigate mediation (n = 9, 81.8%)	- Appropriate statistical analysis (i.e. inappropriate → score 0; appropriate but insufficient information (e.g. whether assumptions were met) → score 1; appropriate → score 2) [24] - Criteria 2 - Mediation test: cumulative score from M1-M5 depending on number of items satisfied (1= coefficient test, 2 = causal steps approach, 3 = testing statistical significance of indirect path, 4 = examining alternative mediators by comparing different indirect effects, 5 = testing alternative order of variables to establish causality) [25] - Were statistically appropriate/ acceptable methods of data analysis used? (This includes the product of coefficient approach with bootstrapped confidence intervals, structural equation modelling, latent growth modelling, and causal mediation analysis) [26]
3. Non-bias-related aspect (n = 10, 90.9%)		
3.1.1	Reporting quality; Power and sample size; Theoritical rationale; Generalizability; Findings and results (n = 10, 90.9%)	- Study reports confidence intervals of mediated effect (CIs for paths a and if Baron and Kenny approach, or CIs for indirect path if Product-Of-Coefficient used) [23] - Did the study report a power calculation and was the study adequately powered to detect mediation?[16,17,20–22] - Did the study cite a theoretical framework?[16,17,19–22] - Inclusion and exclusion criteria – judged on the appropriateness for the aims of this systematic review (i.e. insufficient → score 0; minimally sufficient → score 1; sufficient → score 2)[24] - Was the change in the potential mediator correlated with change in outcome?[20]

appendix 2. One tool (Williams, 2018) was excluded from this analysis as its content was not accessible [15]. Apart from the study design and some non-bias-related aspects (e.g. reporting quality, generalizability of the findings), we identified three mediation-specific bias domains across eleven tools (Table 4). These included (i) temporal order bias (8 tools, 78.7%); (ii) confounding bias (7 tools, 63.6%) and (iii) bias due to the inappropriate use of a statistical approach to investigate mediation (9 tools, 81.8%). Some examples of signalling questions for each bias domain are provided in Table 3.

The domain profiles of the tools were clustered in Fig. 3. The first cluster consisted of three tools (Mansell, 2013 [20]; Cerin, 2009 [17] and Hoppen, 2018 [24]). All

Table 4. Explanations for mediation-specific bias domains identified across 11 different quality/risk of bias assessment tools for mediation analysis

No.	Bias domain	Explanations
1	Temporal order bias	- This bias occurs when data of the mediator, the outcome (and probably the exposure in observational studies) are simultaneously measured at the same time point, overlooking the temporal perspective of a mediational process [27,28].
2	Confounding bias	- Even in the context of a randomized controlled trial, mediation findings can be subject to mediator-outcome confounding bias, as treatment randomization does not ensure that the mediator is free of any systematic relationship to unobserved variables [2,29]. - Apart from mediator-outcome confounding, exposure-mediator and exposure-outcome confounding may also present in mediation analyses using observational data, because participants in observational studies are not randomly allocated to receive the exposure.
3	Bias due to the inappropriate use of a statistical approach to investigate mediation	- Many traditional approaches to mediation (e.g. the Baron and Kenny causal step approach, the Sobel's test, etc.) are no longer appropriate in practice as they are too conservative and/or based on strong distributional assumptions[28,29]. - In many settings (e.g., when the analysis involves non-rare binary or time-to-event endpoints that necessitate nonlinear models), some statistical approaches to mediation (e.g. the counterfactual-based mediation approaches) are deemed more appropriate than other approaches (e.g. the product- and difference-of-coefficient approaches)[2,29].

tools included at least three items in the *non-bias related* domain (30 to 44% of each tool). The second cluster consisted of 2 tools, where one tool (i.e. Rhodes, 2010 [18]) was an update of the other tool (i.e. Luban, 2008 [16]) by adding three more items in the *temporal order bias, non-mediation-specific biases* and *bias due to the inappropriate use of a statistical approach to investigate mediation* domains. The third cluster consisted of three tools, i.e. Windgassen, 2017 [23]; Lee, 2015 [22] and Cortés-García, 2019 [26]. These tools included one item in the *bias due to the inappropriate use of a statistical approach to investigate mediation* domain (11.1 to 14.3% of each tool) and at least three items in the *non-bias related* domain (37.5 to 44.4% of each tool). The forth cluster consisted of two tools (Gu, 2015 [21] and van Stralen, 2011 [19]) which included five items in the *non-mediation-specific biases* domain (31.3 to 50.0% of each tool). The final cluster only consisted of one tool, i.e. Claassen, 2019 [25]. This tool indeed had very different domain profiles compared to other tools in other clusters (Fig. 3).

4. Discussion

In this methodological overview, we updated the previous review by Cashin et al [5] to investigate the quality assessment practice in recently published systematic reviews of mediation studies. We found that many reviews did not assess risk of bias, and those that did often used a quality assessment tool that was not designed to evaluate biases specific to mediation analyses. This is problematic as current mediation analyses are often subject to additional biases that are not captured by risk of bias tools for RCTs and observational studies [29,30].

In fairness to the reviewed articles, many of the mediation-specific quality assessment tools have appeared recently and have not been widely disseminated. These tools, moreover, are not consensus-based and not rigorously developed or validated, with the scope (i.e., what is being addressed) often undefined. As a result, they often include a mix of different domains addressing not only internal validity (i.e., risk of bias) but also external validity (i.e., generalizability), power calculation, sample size, theoretical rationale and reporting quality of the mediation analysis. Many of these tools make use of summary scores to assess the overall quality of the assessed study. However, numerical summary quality scores have been shown to be poor indicators of study quality, and so, alternatives to their use should be encouraged[31–33]. Besides, the identified tools did not provide clear guidance to the users on how to answer each signalling question. For instance, some tools evaluated whether the mediation analysis was implemented by using statistically appropriate/acceptable methods. Assessing this bias is reasonable, as recent developments in the field of mediation has made clear that some classical mediation approaches (such as the product-of-coefficient and the difference-of-coefficient approaches) cannot generally be used for binary and/or time-to-event mediator and outcome data [2]. However, such explanations did not feature in any of the reviewed tools. This lack of explanation and guidance may induce confusion, hence decreasing the validity of risk of bias assessments in practice.

Regarding the content, it is worth noting that the mediation-related biases were not adequately discussed in the identified tools. For instance, the presence of unmeasured mediator-outcome confounding may threaten the validity of mediation findings, even in the context of an RCT

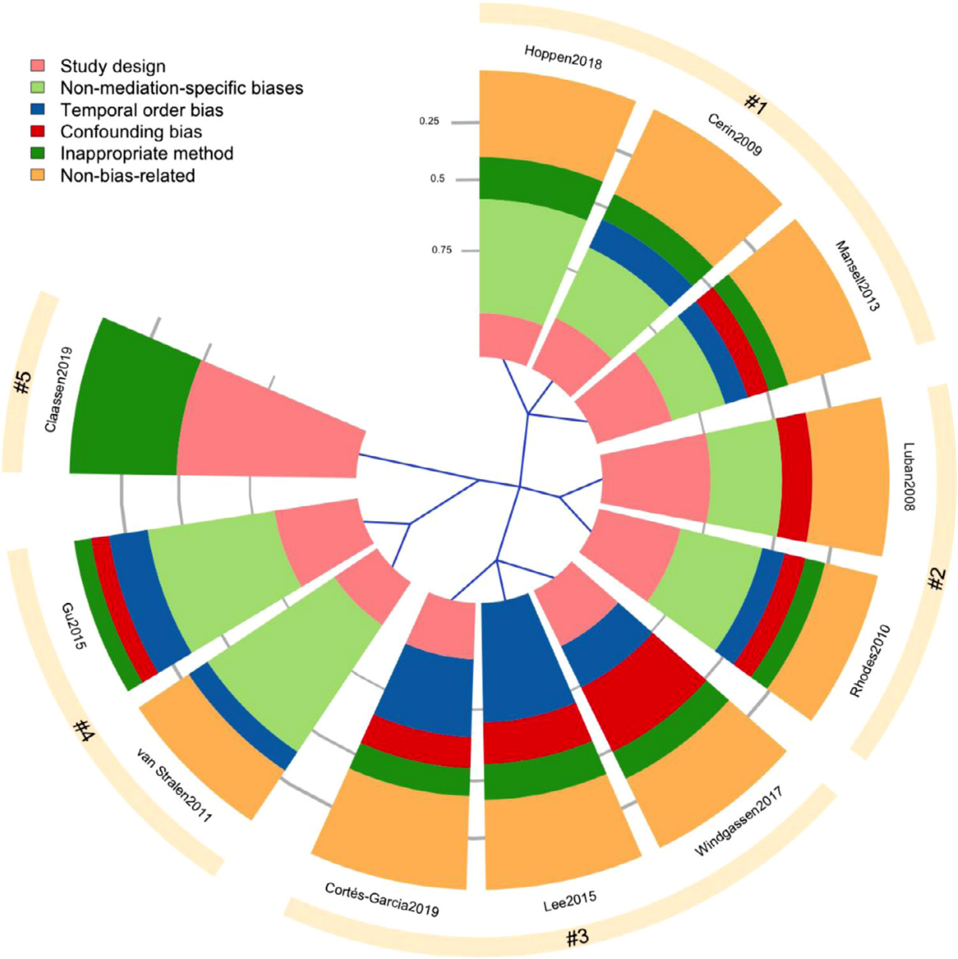

Fig. 3. Hierarchical clustering of 11/12 tools based on the six quality domains. The figure shows which quality domains are present in each tool. One tool (Williams, 2018) was excluded from this analysis as its content was not accessible. A slice of the chart represents a tool. Each slice is divided into different sectors indicating different quality domains. The area of each sector corresponds to the proportion of each domain within the tool. For instance, the tool developed by Claassen et al (2019) consists of two domains: *study design,* and *bias due to the inappropriate use of a statistical approach to investigate mediation* – each encompassing 50% of the tool. The blue lines starting from the centre of the chart define how the tools are divided into the four clusters. Clusters #1, #2 and #3 are sub-nodes of a major node grouping all three, meaning that the tools in these clusters have a similar domain profile compared to the tools in clusters #4 and #5.

[1,4]. However, many of the mediation-specific quality assessment tools did not assess this kind of bias. For the tools which considered bias from unmeasured mediator-outcome confounding, they only required that the baseline values of the mediator(s) and/or outcome were taken into account, or that mediator and outcome variables were assessed for change. No tools explicitly required studies to adjust extensively for mediator-outcome and/or mediator-mediator confounding. Likewise, many tools did not assess temporal order bias (i.e., the lack of temporality in the measurement of the mediator(s) and the outcome). Among tools that assessed this bias, the signaling questions only targeted the simplest setting of single mediation analysis, where the mediators and outcome are not repeatedly measured. These questions, therefore, are not appropriate to assess temporal order bias in more complicated setting such as serial mediation analysis (in which one must also assess whether a mediator M_1 is measured before another mediator M_2 if M_1 is assumed to affect M_2, rather than vice versa).

In view of the above concerns, it is important that a consensus-based quality assessment tool for mediation analysis is constructed in the near future. Such a tool should be rigorously developed and validated to overcome the limitations of other tools currently available in the literature. While the construction of this tool could take time, effort and resources, more hands-on and up-to-date tutorials are needed to guide the clinical and applied researchers in critically appraising results of mediation analyses.

Our study has some limitations. Although we implemented a comprehensive search strategy for systematic reviews of mediation studies, we may have missed eligible reviews and quality assessment tools as we did not consider searching any grey literature. Moreover, only half of the data extraction in this review was double-checked by a second reviewer, which might result in potential mistakes. Finally, we limited the eligibility criteria to reports published only in English. It might be the case that there were non-English systematic reviews of mediation studies that would be eligible. Such a limitation is also common in many medical and methodological systematic reviews.

5. Conclusion

The quality assessment practice in recently published systematic reviews of mediation studies is suboptimal, which increases the risk of mediation-specific biases not being properly evaluated.

To improve the quality and consistency of risk of bias assessments for mediation studies, a consensus-based risk of bias tool is needed. This will be a critically important step towards better quality mediation systematic reviews in the future.

Ethics approval

Not applicable.

Authors' contributions

TTV, AC, HL, IB and SV designed the study. TTV directed the study implementation, including quality assurance and control. TTV, AC, CS and PHTT designed the study's analytic strategy. TTV, AC, CS, PHTT and TBN conducted the literature review. TTV, AC, CS, TBN and PHTT prepared the draft of the paper. IB, DM, TV, HL and SV helped critically revise the paper. All authors read and approved the final manuscript.

Data Availability

The dataset supporting the conclusions of this article is provided in its supplementary file.

Acknowledgements

Not applicable.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jclinepi.2021.12.013.

References

[1] VanderWeele T. Explanation in Causal Inference: Methods for Mediation and Interaction, Oxford, New York: Oxford University Press; 2015. 728 p.

[2] VanderWeele TJ. Mediation analysis: a practitioner's guide. Annu Rev Public Health 2016;37:17–32.

[3] MacKinnon DP, Fairchild AJ, Fritz MS. Mediation Analysis. Annu Rev Psychol 2007;58(1):593–614.

[4] Vo T-T, Vansteelandt S. Challenges in systematic reviews and meta-analyses of mediation analyses. ArXiv210312227 Cs Stat [Internet] 2021. [cited 2021 Jun 15]; Available from: http://arxiv.org/abs/2103.12227 .

[5] Cashin AG, Lee H, Lamb SE, Hopewell S, Mansell G, Williams CM, et al. An overview of systematic reviews found suboptimal reporting and methodological limitations of mediation studies investigating causal mechanisms. J Clin Epidemiol [Internet] 2019. [cited 2019 Apr 4]; Available from: http://www.sciencedirect.com/science/article/pii/S0895435618310011 .

[6] Nguyen TQ, Schmid I, Stuart EA. Clarifying causal mediation analysis for the applied researcher: Defining effects based on what we want to learn. Psychol Methods 2021;26(2):255–71.

[7] Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;366:l4898.

[8] Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016;355:i4919.

[9] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71.

[10] mh329. PROSPERO: International prospective register of systematic reviews — University of Leicester [Internet] 2019. [cited]Available from: https://www2.le.ac.uk/library/find/databases/p/Prospero .

[11] Superchi C, González JA, Solà I, Cobo E, Hren D, Boutron I. Tools used to assess the quality of peer review reports: a methodological systematic review. BMC Med Res Methodol 2019;19(1):48.

[12] Wells G, Shea B, O'Connell D, Peterson J, Welch V, Losos M, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality if nonrandomized studies in meta-analyses [Internet] 2021. [cited]Available from: http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp .

[13] Harbour R, Miller J. A new system for grading recommendations in evidence based guidelines. BMJ 2001;323(7308):334–6.

[14] Sterne JAC, Higgins J, Reeves B. A Cochrane risk of bias assessment tool: For non-randomized studies of interventions (ACROBAT-NRSI) [Internet] 2014. [cited 2021 Nov 20]. Available from: https://hero.epa.gov/hero/index.cfm/reference/details/reference_id/2826337 .

[15] Williams J, Bucci S, Berry K, Varese F. Psychological mediators of the association between childhood adversities and psychosis: A systematic review. Clin Psychol Rev Nov 2018;65:175–96.

[16] Lubans DR, Foster C, Biddle SJH. A review of mediators of behavior in interventions to promote physical activity among children and adolescents. Prev Med Nov 2008;47(5):463–70.

[17] Cerin E, Barnett A, Baranowski T. Testing theories of dietary behavior change in youth using the mediating variable model with intervention programs. J Nutr Educ Behav 2009;41(5):309–18.

[18] Rhodes RE, Pfaeffli LA. Mediators of physical activity behaviour change among adult non-clinical populations: a review update. Int J Behav Nutr Phys Act 2010;7:37.

[19] van Stralen MM, Yildirim M, te Velde SJ, Brug J, van Mechelen W, Chinapaw MJM, et al. What works in school-based energy balance behaviour interventions and what does not? A systematic review of mediating mechanisms. Int J Obes 2011;35(10):1251–65 2005.

[20] Mansell G, Kamper SJ, Kent P. Why and how back pain interventions work: what can we do to find out? Best Pract Res Clin Rheumatol 2013;27(5):685–97.

[21] Gu J, Strauss C, Bond R, Cavanagh K. How do mindfulness-based cognitive therapy and mindfulness-based stress reduction improve mental health and wellbeing? A systematic review and meta-analysis of mediation studies. Clin Psychol Rev 2015;37:1–12.

[22] Lee H, Hübscher M, Moseley GL, Kamper SJ, Traeger AC, Mansell G, et al. How does pain lead to disability? A systematic

review and meta-analysis of mediation studies in people with back and neck pain. Pain 2015;156(6):988–97.

[23] Windgassen S, Moss-Morris R, Chilcot J, Sibelli A, Goldsmith K, Chalder T. The journey between brain and gut: A systematic review of psychological mechanisms of treatment effect in irritable bowel syndrome. Br J Health Psychol 2017;22(4):701–36.

[24] Hoppen TH, Chalder T. Childhood adversity as a transdiagnostic risk factor for affective disorders in adulthood: A systematic review focusing on biopsychosocial moderating and mediating variables. Clin Psychol Rev 2018;65:81–151.

[25] Claassen MA, Klein O, Bratanova B, Claes N, Corneille O. A systematic review of psychosocial explanations for the relationship between socioeconomic status and body mass index. Appetite 2019;132:208–21.

[26] Cortés-García L, Takkouche B, Seoane G, Senra C. Mediators linking insecure attachment to eating symptoms: A systematic review and meta-analysis. PloS One 2019;14(3):e0213099.

[27] Winer ES, Cervone D, Bryant J, McKinney C, Liu RT, Nadorff MR. Distinguishing mediational models and analyses in clinical psychology: atemporal associations do not imply causation. J Clin Psychol 2016;72(9):947–55.

[28] Fairchild AJ, McDaniel HL. Best (but oft-forgotten) practices: mediation analysis12. Am J Clin Nutr 2017;105(6):1259–71.

[29] Vo T-T, Superchi C, Boutron I, Vansteelandt S. The conduct and reporting of mediation analysis in recently published randomized controlled trials: results from a methodological systematic review. J Clin Epidemiol 2020;117:78–88.

[30] Lapointe-Shaw L, Bouck Z, Howell NA, Lange T, Orchanian-Cheff A, Austin PC, et al. Mediation analysis with a time-to-event outcome: a review of use and reporting in healthcare research. BMC Med Res Methodol [Internet] 2018;18. [cited 2019 Apr 4]Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6206666/ .

[31] Whiting P, Wolff R, Mallett S, Simera I, Savović J. A proposed framework for developing quality assessment tools. Syst Rev 2017;6(1):204.

[32] Whiting P, Harbord R, Kleijnen J. No role for quality scores in systematic reviews of diagnostic accuracy studies. BMC Med Res Methodol 2005;5:19.

[33] Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 1999;282(11):1054–60.