Journal of Clinical Epidemiology 101 (2018) 61–72

ORIGINAL ARTICLE

Three risk of bias tools lead to opposite conclusions in observational research synthesis

Josep-Maria Losilla[a], Isabel Oliveras[a], Juan A. Marin-Garcia[b], Jaume Vives[a,*]

[a]Department of Psychobiology and Methodology of Health Sciences, Psychology Faculty, Universitat Autònoma de Barcelona, Carrer de la Fortuna, Edifici B, Bellaterra, Barcelona 08193, Spain

[b]Department of Business Management, School of Industrial Engineering, Universitat Politècnica de València, Dept. Organización de Empresas, Edificio 7D, Camino de Vera s/n, Valencia 46022, Spain

Accepted 25 May 2018; Published online 2 June 2018

Abstract

Objectives: The aim of this study was to assess the agreement and compare the performance of three different instruments in assessing risk of bias (RoB) of comparative cohort studies included in a health psychology meta-analysis.

Study Design and Setting: Three tools were applied to 28 primary studies included in the selected meta-analysis: the Newcastle-Ottawa Scale, quality of cohort studies (Q-Coh), and risk of bias in nonrandomized studies of interventions (ROBINS-I).

Results: Interrater agreement varied greatly from tool to tool. For overall RoB, 75% of the studies were rated as low RoB with the Newcastle-Ottawa Scale, 11% of the studies with Q-Coh, and no study was found to be at low RoB using ROBINS-I. No influence of quality ratings on the meta-analysis results was found for any of the tools.

Conclusion: Assessing RoB using the three tools may lead to opposite conclusions, especially at low and high levels of RoB. Domain-based tools (Q-Coh and ROBINS-I) provide a more comprehensive framework for identifying potential sources of bias, which is essential to improving the quality of future research. Both further guidance on the application of RoB tools and improvements in the reporting of primary studies are necessary. © 2018 Elsevier Inc. All rights reserved.

Keywords: Risk of bias; Methodological quality; Systematic review; Meta-analysis; Domain of bias; Quality tool

1. Introduction

Assessing the methodological quality or risk of bias (RoB) of primary studies is an essential component of any systematic review or meta-analysis [1,2] and should play a relevant role in interpreting the results of the review [3]. Moreover, the inclusion of poor-quality studies in a review may lead to invalid conclusions [3,4]. In fact, the results of such quality assessments often exert an important influence on some decisions made in the review process, such as whether to exclude studies not meeting certain quality standards, to perform sensitivity analyses, to determine the strength of evidence, or to guide recommendations for future research and clinical practice [5,6].

Compared to clinical trials, the quality assessment of observational studies is often more demanding due to the variety of designs comprised and their increased susceptibility to bias [5,7,8]. These difficulties are probably the reason why in some areas such as health psychology, only about half of all reviews that include cohort and case–control studies assessed the RoB of the primary studies [9]. Although a wide range of tools suitable for observational studies have been reviewed by several authors [10–12], there is no consensus on which is the best procedure or tool to assess RoB in observational designs, despite observational studies are usually included in systematic reviews including those of Cochrane [13]. Moreover, most of these tools were poorly developed, and their developers often failed to follow standard methodological procedures or to test their tools' validity and reliability [10,14]. Thus, RoB assessments of a single study using different tools may lead to different conclusions [4,15,16], both in randomized controlled trials [1,14,17] and in observational studies [7,8,18].

Conflict of interest: None.

Funding: This work was supported by the Spanish Ministry of Science and Innovation (grant number: PSI2014-52962-P). I.O. was supported by funding from a predoctoral grant from the Ministry of Education, Culture and Sport of Spanish Government (grant number: FPU14/04514). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the article.

* Corresponding author. Tel.: +34 93 5812331; fax: +34 93 5812001.
E-mail address: Jaume.Vives@uab.cat (J. Vives).

https://doi.org/10.1016/j.jclinepi.2018.05.021

What is New?

Key findings

- Assessing risk of bias (RoB) using the three tools may lead to opposite conclusions, especially at low and high levels of RoB, where most of the studies were rated as low RoB with the Newcastle-Ottawa Scale (NOS), contrary to risk of bias in nonrandomized studies of interventions (ROBINS-I) with which most of the studies were rated as high RoB, while quality of cohort studies (Q-Coh) showed greater variability. Therefore, both the NOS and ROBINS-I showed low capability in grading RoB in observational studies.
- Correlation between Q-Coh and ROBINS-I was good for most of the domains of bias, while correlations between these two tools and the NOS showed poorer agreement. Raters' assessments of the usability of the tools also reveal the similarities between Q-Coh and ROBINS-I.
- The results of subgroup and meta-regression analyses showed no clear association between RoB and combined effect sizes when a meta-analysis is performed.

What this adds to what was known?

- Although this study has found that Q-Coh and ROBINS-I are comprehensive and valid tools compared to the NOS, their reliability needs to be improved.
- This study provides empirical evidence that the NOS assessment of RoB is overly positive.
- To our knowledge, this is the first time that the properties of ROBINS-I have been tested. When applying ROBINS-I, the use of a target trial makes it difficult to discriminate levels of RoB between observational studies and hinders the understanding of some items.

What is the implication and what should change now?

- To improve the reliability of the tools, two conditions must be met: (1) the development of detailed guidance and training in the application of RoB assessment tools and (2) improvements in the reporting of primary studies.
- In the context of systematic reviews and meta-analysis, RoB assessments make it possible to identify weaknesses in research designs and should guide the improvement of the quality of future studies, which is especially relevant to synthesize the results of nonexperimental research.

Meanwhile, the use of scales that provide a single summary score is strongly discouraged [4,15,19] because it involves the weighting of component items, although some of them may be not related to RoB [3,11]. The alternative seems to perform an RoB assessment based on domains [20–23], which is increasingly applied and apparently provides a more structured framework within which to make qualitative decisions on the overall quality of studies and to detect potential sources of bias [16].

The general purpose of this study was to assess the agreement and compare the performance of three different instruments in assessing the RoB of comparative cohort studies included in a meta-analysis related to health psychology. The selected tools were as follows: (1) NOS [24], the most frequently used scale to assess the quality of cohort and case–control studies [9], which provides a summary score; (2) quality of cohort studies (Q-Coh) [21], a specific domain-based tool to assess the RoB of cohort studies with good psychometric properties; and (3) risk of bias in nonrandomized studies of interventions (ROBINS-I) [22], a new domain-based tool proposed by Cochrane, which is intended to assess RoB in nonrandomized studies of interventions but is also applicable to a wide variety of observational designs [25]. To be more precise, the specific objectives are as follows:

- To estimate, for each tool, the degree of interrater agreement when examining items, domains of RoB, and overall quality rating.
- To estimate the level of agreement between tools for specific biases, domains of RoB, and overall quality rating.
- To appraise the qualitative aspects of the tools related to their usability: the average time spent, clarity of instructions and items, coverage, and validity.
- To determine the effect of quality ratings on the results of a meta-analysis.

2. Methods

2.1. Risk of bias assessment tools

The NOS [24] was developed to assess the quality of observational studies included in systematic reviews. This tool exists in separate versions for cohort and case–control designs, although only the scale for cohort studies was applied here. Studies are assessed using eight items broken down into three dimensions: selection (four items), comparability (one item), and exposure for case–control studies or outcome for cohort studies (three items). A study can be awarded a maximum of nine stars. Although the tool's developers have said that the validity and reliability of the tool have been established, no further specific information has been published. Nevertheless, subsequent studies that have tested the NOS have come to quite varied results in terms of interrater reliability and validity [7,26–28].

Q-Coh [21] is a bias domain–based tool specifically intended to assess the RoB of cohort studies and developed by two of the authors of the present study. A pilot of the second version of the tool was applied in this study (Appendix A). This version of Q-Coh is structured around four domains of bias: (1) selection (three items), (2) confounding (four items), (3) exposure measures (three items), and (4) outcome measures (five items). Each domain of RoB is evaluated as "yes" (potential bias) and "no" (no bias), and the overall assessment of RoB based on the four domains is rated as "low RoB," "moderate RoB," or "high RoB." The tool also includes several previous considerations about the most important confounding factors in the research field under study, the acceptable percentage of missing data, and the exposure and outcome variables of interest. The reliability and validity of the original tool were established by the developers, with interrater agreement kappa values ranging from 0.60 to 0.87 for the different domains (0.75 for overall assessment) and a weighted kappa value equal to 0.41 ($P = 0.003$) for validity analyzed by studying the agreement of the ratings of the overall RoB of the studies with an external rating.

The ROBINS-I tool [22] is a new published tool from Cochrane for assessing RoB in non-randomized studies of the effects of interventions. The authors define bias as "systematic difference between the results of the non-randomized studies of the effects of interventions and the results expected from the target trial" [22], where the "target trial" is a hypothetical randomized trial without threat of bias. Although ROBINS-I was designed to compare the effects of two or more interventions, the term "intervention" refers here to either treatment or exposure, thus including studies in which no intervention was carried out by the investigators [25]. The tool is structured around seven domains of bias: (1) bias due to confounding, (2) bias in selection of participants for the study, (3) bias in classification of interventions, (4) bias due to deviations from intended interventions, (5) bias due to missing data, (6) bias in measurement of outcomes, and (7) bias in selection of the reported results. Every domain includes a series of signaling questions to help the reviewers judge the RoB in the same four categories as in the overall RoB (low, moderate, serious, and critical RoB), based on the responses given to the signaling questions. The developers also provide detailed guidance on the use of ROBINS-I [22], but, to the best of our knowledge, to date, there are no published data on the reliability and validity of this tool.

2.2. Selected studies

A systematic review with meta-analysis [29] in the field of health psychology, including 28 comparative cohort studies (the study design that is common to the three tools), was selected to test the properties of the tools. The references of the 28 primary studies are provided in Appendix B. The selected review explores the prospective association between depression and the risk of developing stroke in adults. No intervention was applied to the participants of the studies. The main meta-analysis, including 31 effect sizes, was performed using a random-effects model, and its summary effect shows an increased risk of stroke morbidity and mortality in depressed participants compared with nondepressed individuals, with a hazard ratio (HR) estimate of 1.45 (95% CI, 1.29–1.63) and moderate to high heterogeneity ($I^2 = 66.0\%$; Cochrane Q statistic = 88.1, $P < 0.001$).

2.3. Procedure

The three tools were pilot-tested by two independent raters (I.O. and J.A.M.-G.) on a study not included in the meta-analysis. The study team discussed and agreed upon some previous considerations that were necessary for the application of the tools (ie, exposure and outcome of interest, the most important confounding factors, acceptable percentage of missing data, and minimum duration of follow-up). A randomization scheme was then developed to randomize the order of studies and the order of the tools applied by each rater. Following this randomization scheme, the two raters independently applied each of the three tools to the 28 selected papers over the same time frame (approximately 2 months) to reduce both intrarater and interrater variations in quality assessment.

Consensus on quality ratings was reached via the following process: (1) if the scores of both raters were the same, then that score was used as a consensus score; (2) if the scores differed, a consensus score was agreed upon by both raters; and (3) if the raters were unable to reach a consensus, then the score provided by a third rater (J.-M.L.) was used for discussion and final agreement.

Finally, the ease of use of the tools was rated on a five-point scale for each of the following aspects (Appendix C): (1) coverage of the tool (the extent to which the tool covers all the relevant domains of bias and features or if it includes nonrelevant items), (2) clarity of instructions, (3) clarity of the items, and (4) ability to discriminate studies of high RoB from low RoB.

2.4. Analysis

For the purposes of comparison and using a strategy similar to that applied by other authors (eg, [30–33]), three categories were created using the NOS scores: studies scoring 0–3, 4–6, and 7–9 were deemed to be of low, moderate, and high quality, respectively. When at least one item in an NOS domain reflected bias, RoB was considered to be present in that domain.

Traditionally, the most commonly used statistics to evaluate interrater agreement are Cohen's kappa coefficient

[34] or its generalization for multiple raters proposed by Fleiss [35]. However, when the prevalence of a given response category is notably high or low, these statistics are not advisable. Under these circumstances, the "kappa paradox" appears, meaning that the value of the kappa statistic is low even when the observed proportion of agreement is significantly high [36]. A second kappa paradox arises when the extent to which raters disagree on the proportion of cases in each response category is large, resulting in kappa values higher than when this bias is low or absent. Given that kappa is difficult to interpret in the presence of different prevalence or bias, several studies have recommended including detailed information about the proportions of specific agreement between raters for each response category to make it possible to evaluate the possible effects of prevalence or bias [37—39]. In addition, in the presence of different prevalence or bias, a widely used alternative to Cohen's kappa is the Prevalence-Adjusted and Bias-Adjusted Kappa (PABAK) proposed by Byrt et al. [40].

Therefore, several statistics are provided for each item, domain, and overall RoB: the proportion of interrater agreement, the proportion of choices for each response category, and the Fleiss kappa statistic (or PABAK when necessary). Following the recommendations of Q-Coh and ROBINS-I authors, interrater reliability was calculated once the equivalent categories of response had been merged (ie, yes/probably yes, no/probably no for ROBINS-I, yes/presumably, and yes/not necessary for Q-Coh). Landis and Koch's [41] criteria were used to define the levels of agreement for kappa (or PABAK): no agreement (<0), slight (0—0.20), fair (0.21—0.40), moderate (0.41—0.60), substantial (0.61—0.80), and almost perfect agreement (0.81—1). All these analyses were performed using the irr package (v.0.84) [42] for R version 3.3.2 (R Foundation for Statistical Computing, Vienna, Austria).

To determine agreement between tools, the items of each tool were classified into common domains and aspects of bias. The comparison of the groups of related items was performed applying the nonparametric Kendall's tau-b correlation coefficient [43], using the Statistical Package for Social Sciences (SPSS) version 20.0 (SPSS, Inc, 2009, Chicago, IL, USA). Good agreement between the tools indicates that they estimate comparable constructs, offering an indirect measure of concurrent validity.

Beyond the qualitative assessment of the tools, some indicators of their usability were also calculated using Microsoft Excel 2016. The median time spent on the application of the tools was calculated, counting only the scoring time in the cases where each tool was applied first. Each tool was also measured in terms of the mode number of items in which the answer was "not reported" or "no information" and the mode number of items that required a third rater to reach a consensus.

Finally, subgroup, meta-regression, and sensitivity analyses were performed to test the effects of quality ratings on the results of the meta-analysis for every tool, both in terms of each individual domain of bias and of overall RoB, following the classification previously created for the comparison between tools. These analyses were performed using Comprehensive Meta-analysis version 3 (BioStat, Inc, 2014, Englewood, NJ, USA), following a mixed-effects model for subgroup analyses and a random-effects model with the Knapp Hartung adjustment for meta-regressions [44].

3. Results

3.1. Risk of bias assessment

Fig. 1 shows a summary of the consensus results of RoB assessment for each tool, for overall RoB and by domain. The NOS scores ranged from 5 to 9, with a median and mode of 8 (25th percentile [p25] = 6; p75 = 9). Once the studies were classified into categories, there were 21 studies with low RoB and seven studies with moderate RoB. None of the studies were placed in the category of high RoB. According to the Q-Coh results, three studies were classified as low RoB, 11 studies as moderate RoB, and 14 studies as high RoB. Finally, when ROBINS-I was applied, four studies were deemed to have moderate RoB and 24 to have serious RoB, with no studies classified as having critical RoB or no information categories. The fourth domain of RoB in ROBINS-I (bias due to deviations from intended interventions) was judged inapplicable due to the nonexperimental design of the studies.

3.2. Interrater reliability

Table 1 presents the proportion of agreement between raters for items (all the tools), individual domains, and overall RoB (Q-Coh and ROBINS-I) and the kappa (or PABAK) values, once equivalent response categories were joined as specified in Section 2. It was not possible to synthesize Q-Coh results for items corresponding to the confusion domain (items from 4 to 7) because they were applied to all individual confounding variables within the same study.

Interrater agreement for the NOS items ranged from fair to almost perfect (kappa/PABAK: 0.20—1), with a proportion of agreement between raters ranging from 46% to 100%. Agreement between raters for Q-Coh ranged from no agreement to almost perfect agreement (kappa/PABAK: −0.05, 0.93) for items and no agreement to substantial agreement (kappa/PABAK: −0.20 to 0.64) for domains, registering no agreement for overall RoB (kappa = −0.16). The percentage of agreement ranged from 39% to 96% for items and from 50% to 82% for domains, and the figure was 25% for overall RoB. The results for ROBINS-I ranged between slight and almost perfect agreement for items (kappa/PABAK: 0.00 to 1) and domains (kappa/PABAK: 0.08 to 0.93) and showed moderate agreement for overall RoB (PABAK = 0.57). The proportion

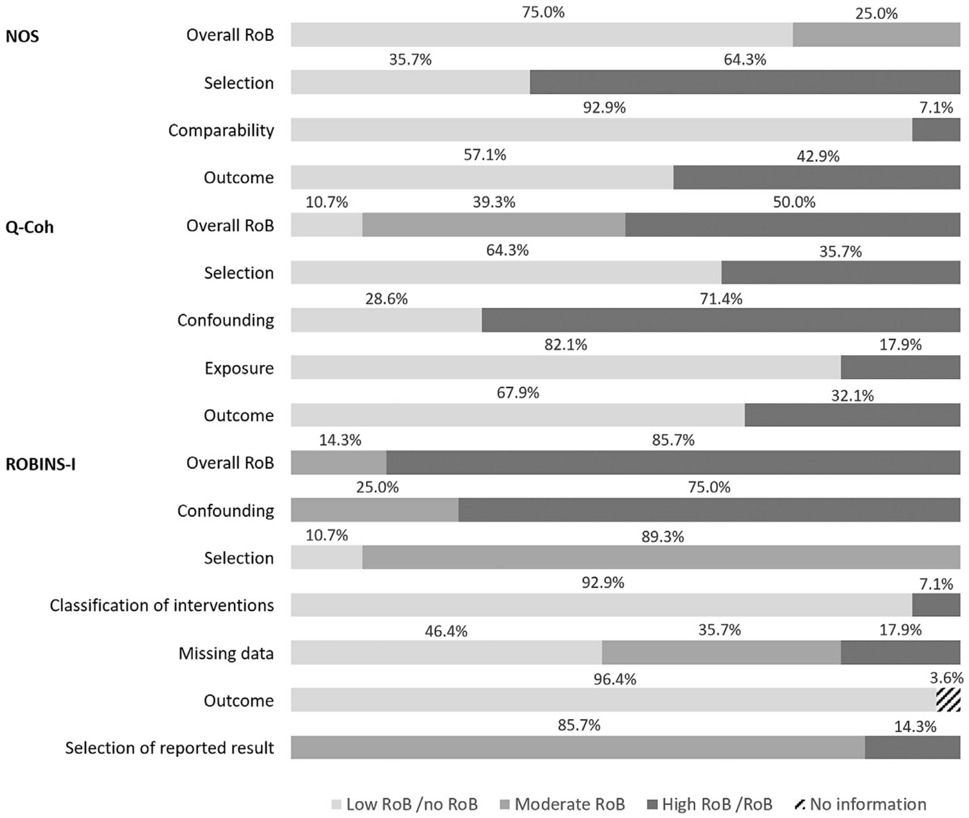

Fig. 1. Consensus results of risk of bias assessment for overall risk of bias and domains of bias for each tool. Categories for overall risk of bias were as follows: low, moderate, and high risk of bias for NOS and Q-Coh; low, moderate, serious, critical risk of bias, and no information for ROBINS-I. Categories for domains of bias were as follows: risk of bias "yes" or "no" for NOS and Q-Coh; for ROBINS-I, the same categories as overall risk of bias were applied. NOS, Newcastle-Ottawa Scale; Q-Coh, quality of cohort studies; RoB, risk of bias; ROBINS-I, risk of bias in nonrandomized studies of interventions.

of agreement between raters ranged from 39% to 100% for items, from 36% to 96% for domains, and was 79% for overall RoB. Note that the lowest agreement values were those corresponding to items or domains related to missing data for Q-Coh and ROBINS-I.

In addition (see Appendix D), over 90% of the responses of the raters were concentrated in a single response category in three items of the NOS (two of them corresponding to comparability domain), only in one item of Q-Coh regarding outcome domain, and finally, in 18 items of ROBINS-I belonging to domains of confounding, selection, intervention, outcome, and reported results. The results for the outcome domain for ROBINS-I also showed a significant lack of variability as 95% of the responses were concentrated in a single category.

3.3. Agreement between tools

To make comparisons between the tools, the first step needed was an analysis of the content of the tools and classify related items into 19 aspects of bias, all while preserving the original domains of each of the tools (Table 2). Items related to missing data are distributed among several domains; these items were also grouped together in a single domain and compared.

The agreement coefficients between tools for overall RoB and the different domains are shown in Table 3. Regarding the NOS, only the associations of the missing data domain for Q-Coh and ROBINS-I and the outcome domain for Q-Coh were statistically significant. By contrast, the correlations between Q-Coh and ROBINS-I were significant for overall RoB and for most of the domains, with the exceptions of the selection and outcome domains.

3.4. Usability of the tools

The raters gave generally poor scores to the NOS for coverage and discriminative ability, while the results for the clarity of instructions were very good and the scores for the clarity of the items were moderate to very good (Table 4). The Q-Coh tool obtained moderate to very good rates for all the aspects assessed. The ratings of ROBINS-I were very poor for all the aspects, except for the coverage of the tool, which was considered moderate to good.

Table 4 also shows the median of time spent to rate the studies for each tool. Because the procedure used by the raters consisted of a prior identification of the necessary data for the application of the three RoB

Table 1. Results of interrater agreement

NOS			Q-Coh			ROBINS-I[a]		
Items	P. Overall agreement	Kappa/PABAK	Items and domains	P. Overall agreement	Kappa/PABAK	Items and domains	P. Overall agreement	Kappa/PABAK
Selection 1	0.46	.24	Item 1	0.79	0.57[b]	1.1	1.00	1.00[b]
Selection 2	0.96	.93[b]	Item 2	0.96	0.93[b]	1.2	0.93	0.86[b]
Selection 3	0.96	.93[b]	Item 3	0.46	0.23	1.3	0.93	0.86[b]
Selection 4	0.75	.50[b]	Selection	0.61	0.32	1.4	0.79	0.57[b]
Comparability a	0.93	.86[b]	Confounding[c]	0.82	0.64[b]	1.5	0.68	0.36[b]
Comparability b	1.00	1.00[b]	Item 8	0.79	0.57[b]	1.6	0.96	0.93[b]
Outcome 1	0.68	.36[b]	Item 9	0.64	0.29[b]	1.7	1.00	1.00[b]
Outcome 2	0.96	.93[b]	Item 10	0.39	0.02	1.8	1.00	1.00[b]
Outcome 3	0.61	.21[b]	Exposure	0.50	−0.20	Confounding	0.75	0.50[b]
			Item 11	0.71	0.43[b]	2.1	0.82	0.64[b]
			Item 12	0.89	0.79[b]	2.2	0.82	0.64[b]
			Item 13	0.93	0.86[b]	2.3	0.82	0.64[b]
			Item 14	0.79	0.57[b]	2.4	0.96	0.93[b]
			Item 15	0.39	−0.05	2.5	0.89	0.79[b]
			Outcome	0.64	0.29[b]	Selection	0.82	0.64[b]
			Overall RoB	0.25	−0.16	3.1	0.82	0.64[b]
						3.2	0.96	0.93[b]
						3.3	1.00	1.00[b]
						Intervention	0.79	0.57[b]
						5.1	0.75	0.50[b]
						5.2	0.50	0.14
						5.3	0.43	0.00
						5.4	0.39	0.11
						5.5	0.54	0.08
						Missing data	0.36	0.08
						6.1	1.00	1.00[b]
						6.2	0.86	0.71[b]
						6.3	1.00	1.00
						6.4	1.00	1.00
						Outcome	0.96	0.93[b]
						7.1	0.75	0.50[b]
						7.2	0.57	0.14[b]
						7.3	0.89	0.79[b]
						Reported result	0.64	0.29[b]
						Overall RoB	0.79	0.57[b]

Abbreviations: NOS, Newcastle-Ottawa Scale; P. Overall agreement, proportion of agreement between raters; PABAK, prevalence-adjusted and bias-adjusted Kappa; RoB, risk of bias; ROBINS-I, risk of bias in nonrandomized studies of interventions; Q-Coh, quality of cohort studies.

[a] Domain 4 of ROBINS-I ''Bias due to deviations from intended interventions'' was considered not applicable.

[b] PABAK.

[c] Agreement for items of confounding domain (4—7) could not be calculated.

tools, the time spent in data extraction could not be accounted for individually for each tool. However, it must be pointed out that the time invested in ROBINS-I training was significantly higher compared to Q-Coh, while the time spent in the NOS was significantly lower compared to both. Results as to the number of items that were answered with a ''not reported'' or ''no information'' option, as well as items that required a third-rater judgment to achieve consensus, highlighted the similarities between Q-Coh and ROBINS-I, as well as the differences between the NOS and the other two tools.

Table 2. Items from the three tools classified into common domains and aspects of bias

Domains and aspects of bias	NOS items	Q-Coh items	ROBINS-I signaling questions
Confounding/comparability			
Potential of confounding			1.1
Baseline confounding factors	C1a, C1b	4, 7	1.4, 1.5, 1.6
Confounding during follow-up		4, 7	1.2, 1.3, 1.7, 1.8
Missing data on confounders		5, 6	5.3
Selection			
Representativeness	S1		
Exclusion of participants using different criteria	S2	3	
Selection based on variables after start			2.1, 2.2, 2.3, 2.5
Outcome not present at start	S4	1	
Coincidence of intervention and follow-up start		2	2.4, 2.5
Exposure			
Exposure measure	S3	8	
Classification of participants			3.1, 3.2, 3.3
Missing data on exposure		9, 10	5.2
Outcome			
Blinding of assessors		12	6.1, 6.2
Outcome measure	O1	11	6.3, 6.4
Length of follow-up	O2	13	
Attrition/lost to follow-up	O3	14, 15	5.1
Missing data	O3	5, 6, 9, 10, 14, 15	5.1, 5.2, 5.3, 5.4, 5.5
Selective reporting of results			7.1, 7.2, 7.3

Abbreviations: NOS, Newcastle-Ottawa Scale; ROBINS-I, risk of bias in nonrandomized studies of interventions; Q-Coh, quality of cohort studies.

Additional raters' comments were centered in the poor quality of reporting of most primary studies, especially regarding data loss, and in the difficulties involved in applying the ROBINS-I tool to nonexperimental studies. Another comment was on the need to come to a more detailed agreement on the criteria to be used before the application of the tools, thus allowing the raters to make quality assessment decisions with greater confidence.

Table 3. Results of agreement between tools

Domains of bias	Tools	NOS	Q-Coh
Overall risk of bias	ROBINS-I	−0.058	0.580[a]
	NOS	-	−0.200
Confounding/comparability	ROBINS-I	−0.160	−0.913[a]
	NOS	-	0.175
Selection	ROBINS-I	0.250	−0.258
	NOS	-	0.167
Exposure/classification of interventions	ROBINS-I	0.000	−0.595[b]
	NOS	-	0.093
Outcome	ROBINS-I	0.167	0.132
	NOS	-	0.640[a]
Missing data	ROBINS-I	−0.546[a]	−0.691[a]
	NOS	-	0.683[a]

Abbreviations: NOS, Newcastle-Ottawa Scale; ROBINS-I, risk of bias in nonrandomized studies of interventions; Q-Coh, quality of cohort studies.

Kendall's tau-b correlation coefficient and its significance test were used.

[a] $P < 0.01$.

[b] $P < 0.05$.

3.5. *Effect of quality rating on meta-analysis results*

Only the subgroup analysis for outcome domain of ROBINS-I ($P < 0.001$) and the meta-regression analyses for selection and outcome domains of ROBINS-I ($P = 0.023$ and $P = 0.001$, respectively) led to statistically significant results (Appendix E). However, these results cannot be considered relevant because of the low number of studies included in some of the analysis categories. In any case, it is worth commenting on some trends observed in the results of these analyses.

Fig. 2 shows the results of the subgroup analyses for each tool and each domain of bias and for overall RoB. Equivalent levels of RoB of each tool were grouped and classified by domains to facilitate comparison between tools. No significant differences were found between the tools for overall RoB. Moreover, no association between the level of bias and estimated effect sizes was found for any of the tools. Nevertheless, despite the absence of statistically significant results, the confounding domain in both Q-Coh and ROBINS-I shows a trend in the sense that the smaller the bias the smaller the effect size. The same does not happen with the NOS. It should also be noted that in the

Table 4. Summary of the usability of the tools

Attribute	NOS	Q-Coh	ROBINS-I
Coverage of the tool[a]	2–3	4–5	2–3
Clarity of instructions[a]	5	3–4	1
Clarity of items[a]	3–5	4	1
Discriminative ability[a]	2	3–4	1
Time[b]	4 (2, 6)	13 (11, 20)	17 (14, 20)
Number of items answered with NR/NI option[c]	0 (0–1)	2 (0–6)	1 (1–7)
Number of items requiring a third rater for consensus[c]	0 (0–1)	0 (0–6)	0 (0–10)

Abbreviations: NR, not reported; NI, no information; NOS, Newcastle-Ottawa Scale; ROBINS-I, risk of bias in nonrandomized studies of interventions; Q-Coh, quality of cohort studies.

[a] Range of scores. Each of these attributes was rated from 1 (poor) to 5 (excellent).

[b] Median of minutes in the cases where each tool was first applied (25th and 75th percentiles in parentheses). Only accounted for scoring time. Time spent identifying relevant information before data extraction varies considerably depending on the tool.

[c] Mode number (range in parentheses).

exposure and response domains, the results are more homogeneous for the three tools, with the trend in the opposite direction to that of the confounding domain (ie, the lower the RoB, the greater the effect size). Finally, sensitivity analyses excluding the studies at high RoB also provided nonrelevant results (NOS HR 95% CI: 1.29–1.63; Q-Coh HR 95% CI: 1.19–1.90; ROBINS-I HR 95% CI: 0.93–2.02). However, the number of studies included showed that the least demanding tool was the NOS ($n = 31$; ie, all studies) compared to Q-Coh ($n = 15$) and ROBINS-I ($n = 4$).

4. Discussion

Our comparison of three tools for RoB assessment of nonexperimental studies suggests that we are dealing here with three different approaches to RoB assessment, each of which could lead to different conclusions about the final quality grade assigned to each study. In this study, no agreement between tools was found for overall RoB. While 75% of the studies can be considered to be at low RoB when the NOS is applied, 86% of the studies would be at serious RoB according to ROBINS-I. Overall RoB measured with Q-Coh showed greater variability (11% low, 39% moderate, and 50% high RoB). This lack of agreement corroborates the findings of a great deal of the previous work comparing quality tools for both experimental and nonexperimental studies [1,4,14,17].

The findings on interrater agreement of the NOS are consistent with those of Hootman et al. [7], showing moderate to good interrater reliability and good usability (ie, clarity of items, short scoring time, and ease of consensus). By contrast, the lesser degree of agreement between raters found in Q-Coh and ROBINS-I can be attributed to the broader scope of these tools, which implies a more comprehensive analysis of the primary studies than the NOS requires. Furthermore, Q-Coh and ROBINS-I are more demanding in terms of the amount of information collected and the level of detail used in assessing RoB, but, unfortunately, the quality of reporting of primary studies was not always up to the standards set by these demands. In fact, among the main causes that have been pointed out in the literature to explain low interrater agreement is the difficulty in extracting certain specific information from poorly reported studies [1,14]. The fact that Q-Coh and ROBINS-I obtained the lowest agreements in items related to missing data would also point in this direction. It should be noted that, although agreement between raters in RoB domains for ROBINS-I and Q-Coh has a similar range, it is not the case of overall RoB, where ROBINS-I offers better results compared to Q-Coh. This might be explained by the detailed algorithm for the overall RoB judgment in ROBINS-I, which leaves little margin to rater's decision.

On the other hand, in our opinion, some items are hard to understand and may have negatively affected the interrater agreement. This could be due to the fact that ROBINS-I identifies a target trial as the gold standard against which all observational studies are assessed, as well as the fact that ROBINS-I "to keep the analogy with the target trial (…) uses the term 'intervention' groups to refer to 'treatment' or 'exposure' groups in observational studies even though in such studies no actual intervention was implemented by the investigators" [p.4, 21]. Furthermore, the use in ROBINS-I of a target trial as a reference also makes it difficult to discriminate between the distinct levels of RoB in different observational studies. For example, compared to a target trial, no observational study can achieve low RoB in the confusion domain meaning that no observational study can be given the grade of low overall RoB. In this context, the difficulties we found in understanding the ROBINS-I items may have led to an overagreement in the previous phase of the application of the tool, as discussed later in our study limitations.

Regarding agreement between the tools, Q-Coh and ROBINS-I showed good correlation for overall RoB and for three of five domains of bias, whereas correlations between these two tools and the NOS showed poorer agreement. In this sense, these results are within the realm of what is to be expected, especially if we consider that Q-Coh and ROBINS are tools based on bias domains, while the NOS is a global scoring scale. Only the selection domain showed no significant correlation between any of the tools, which is probably due to the different definition and conceptualization of this domain in the three tools. Otherwise, it is somewhat surprising that no significant correlation was found in the outcome domain for ROBINS-I.

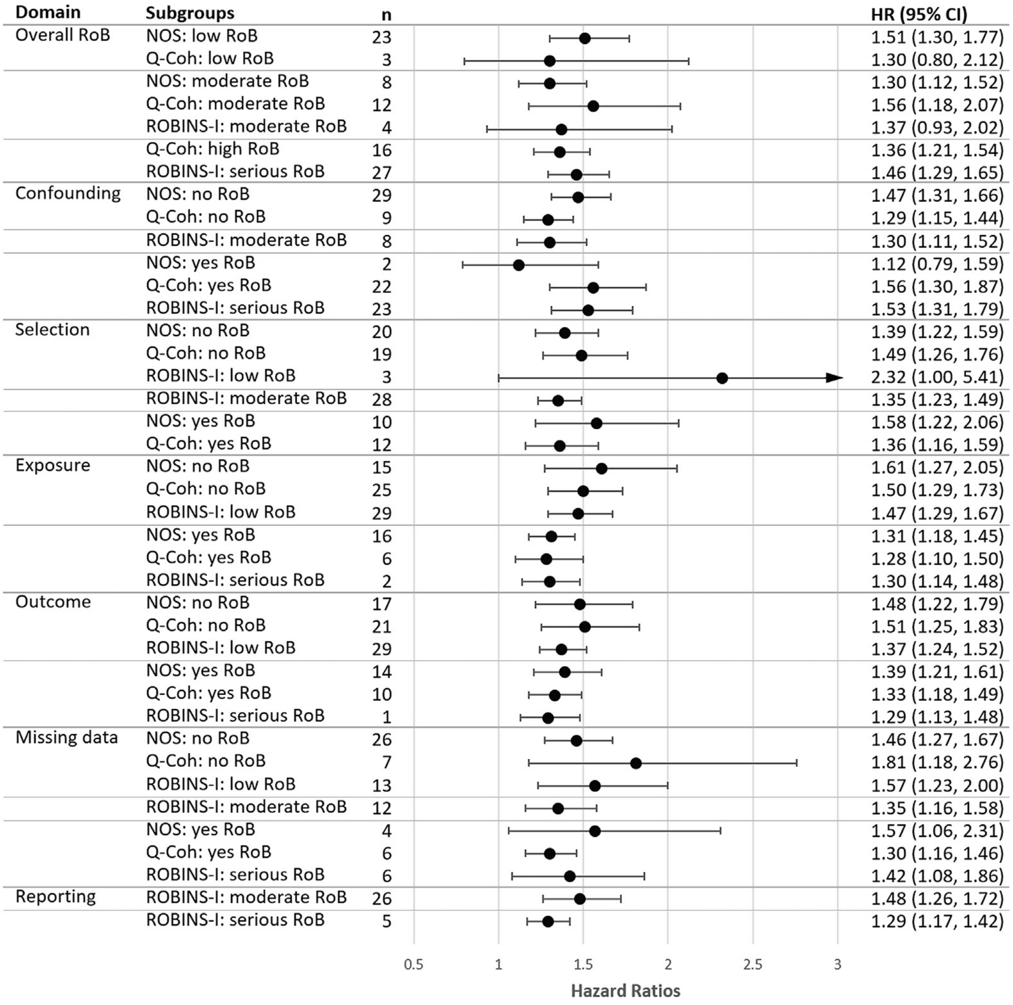

Fig. 2. Forest plot of subgroup analyses results for the Newcastle-Ottawa Scale (NOS), quality of cohort studies (Q-Coh), and risk of bias in non-randomized studies of interventions group. RoB, risk of bias.

This discordance of ROBINS-I could be explained by the lack of direct assessment of the validity and reliability of data collection methods.

The raters' assessments of the usability of the tools reveal the similarities between Q-Coh and ROBINS-I, especially regarding the coverage of the tool, scoring time, loss of information, and ease of consensus. However, the clarity of instructions and items of ROBINS-I, as well as its discriminative ability, were given very low scores by both raters. In addition, as expected, both tools clearly differ from the NOS in most aspects of usability evaluated. In this sense, the shorter scoring time required by the NOS may be one of the reasons for its greater generalization of use.

Finally, although the results of subgroup and meta-regression analyses showed some clear trends when Q-Coh or ROBINS-I are applied, almost all the estimates were small and nonsignificant. Our results are consistent with some previous literature studies that had found no association between quality rating and combined effect sizes [45−47]. Although other studies reported significant effects [16,48−50], there seems to be no clear patterns of associations [51,52]. It might be the case that low variability in RoB is hindering the emergence of an association between RoB and effect size estimates because only moderate- to high-quality studies tend to be included in meta-analyses.

4.1. Strengths and limitations

To the best of our knowledge, this is the first time that the reliability and validity of ROBINS-I have been tested. Furthermore, this is the first study to our knowledge to compare the performance of two domain-based tools and a composite scoring scale applied to observational research.

However, these findings are subject to several limitations. First, the confusing instructions of ROBINS-I and its use of a target trial raised serious doubts among the raters during the pilot phase. Having a trial as reference study forced us to agree to very specific criteria not covered by the tool itself. These specific criteria allowed us to apply the tool in a noninterventional context, where ROBINS-I is hardly applicable. We expect ROBINS-E [53], currently under development, to overcome this shortcoming. Unlike the NOS and Q-Coh, this pre-agreement in ROBINS-I entailed an ad hoc tailoring for which this tool provides no guidelines. This nonstandardized adaptation of ROBINS-I could have been properly evaluated if two teams of raters had been included. Second, there is no gold standard to adequately test concurrent validity, although good correlations between Q-Coh and ROBINS-I indicate that they are measuring similar constructs. Third, the relatively small number of studies considered in some categories of RoB limited the power of subgroup analyses and meta-regressions, leading to wide confidence intervals in those subgroups with few studies. Finally, RoB categories for the NOS were obtained from the overall quantitative score by setting cutoff points, which may be somewhat arbitrary.

4.2. Recommendations and future research

There are some questions arising from our findings, which should be explored. First, it is not clear how reviewers should handle quality assessment in observational research and whether they should take as a reference a target trial or should assess studies against the best available evidence [8]. Although it seems that the tendency is to choose the first option [25,54,55], this seems to be in detriment of the ability to discriminate different levels of quality when only observational studies are being assessed.

Moreover, although it seems domain-based tools showed better attributes and properties than composite scores [16], it is essential to find new methods or procedures that allow for improving the reliability of these tools. This improvement seems to depend on two essential conditions: detailed guidance and training in applying RoB assessment tools, and clear and complete reporting of primary studies. Specific guidance for RoB tools should include clear decision rules to reduce the sort of discrepancies that arose from differing interpretations of the tool [47]. Moreover, Faggion [56] suggested that researchers have to make accessible the rationale used for supporting their judgments to the end users of systematic reviews. Regarding the quality of the reporting, it has proven to be crucial for carrying out a proper RoB assessment [14,47,48]. In our experience, it is too often difficult or even impossible to gather the necessary information to assess certain domains of bias (eg, missing data). This situation is likely to improve considerably if scientific journals systematically include the results of the implementation of reporting guidelines in its publications, as some journals have already done (eg, BJU International [57]).

5. Conclusions

The present study, comparing the performance of three different tools when assessing the RoB of 28 cohort studies, shows that assessing RoB on the same study using different tools may lead to opposite conclusions, especially at low and high levels of RoB, where most of the studies were rated as low RoB with the NOS, contrary to ROBINS-I with which most of the studies were rated as high RoB. Therefore, both the NOS and ROBINS-I showed low capability in grading RoB in observational studies. Our results showed also lower interrater agreement for the most comprehensive tools (Q-Coh and ROBINS-I), as well as lack of association between RoB and combined effect sizes when a meta-analysis is performed.

In light of the results found, we must emphasize the important role of RoB assessment in systematic reviews and in the context of meta-analyses. In this context, RoB assessment provides invaluable information to describe the strength of the evidence found, beyond the usual tests of association between the levels of RoB and the effect estimates of primary studies as potential explanation for part of the observed heterogeneity. The analysis of the results of RoB assessment makes it possible to identify weaknesses in research designs and the most common deficiencies in reporting. This information plays an essential role in guiding the improvement of the quality of studies in a research area, which in turn is a basic objective of research synthesis, especially in nonexperimental research.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2018.05.021.

References

[1] Hartling L, Ospina M, Liang Y, Dryden DM, Hooton N, Krebs Seida J, et al. Risk of bias versus quality assessment of randomised controlled trials: cross sectional study. BMJ 2009;339:b4012.

[2] Johnson BT, Low RE, MacDonald HV. Panning for the gold in health research: incorporating studies' methodological quality in meta-analysis. Psychol Health 2014;30:135–52.

[3] Higgins J, Green S. Cochrane handbook for systematic reviews of interventions version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available at: http://training.cochrane.org/handbook. Accessed December 12, 2017.

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

[5] Centre for Reviews and Dissemination. Systematic reviews: CRD's guidance for undertaking reviews in health care. CRD, University of York; 2009. Available at: https://www.york.ac.uk/crd/guidance/. Accessed May 30, 2017.

[6] Jüni P, Altman DG, Egger M. Systematic reviews in health care—assessing the quality of controlled clinical trials. Br Med J 2001;323:42–6.

[7] Hootman JM, Driban JB, Sitler MR, Harris KP, Cattano NM. Reliability and validity of three quality rating instruments for systematic reviews of observational studies. Res Synth Methods 2011;2:110–8.

[8] Margulis A, Pladevall M. Quality assessment of observational studies in a drug-safety systematic review, comparison of two tools: the Newcastle–Ottawa scale and the RTI item bank. Clin Epidemiol 2014;359–68.

[9] Oliveras I, Losilla J-M, Vives J. Methodological quality is underrated in systematic reviews and meta-analyses in health psychology. J Clin Epidemiol 2017;86:59−70.

[10] Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, et al. Evaluating non-randomised intervention studies. Health Technol Assess 2003;7. iii−x, 1−173.

[11] Sanderson S, Tatt ID, Higgins JPT. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. Int J Epidemiol 2007;36:666−76.

[12] Jarde A, Losilla JM, Vives J. Methodological quality assessment tools of non-experimental studies: a systematic review. An Psicol 2012;28:617−28.

[13] Ijaz S, Verbeek JH, Mischke C, Ruotsalainen J. Inclusion of non-randomized studies in Cochrane systematic reviews was found to be in need of improvement. J Clin Epidemiol 2014;67:645−53.

[14] Armijo-Olivo S, Stiles CR, Hagen NA, Biondo PD, Cummings GG. Assessment of study quality for systematic reviews: a comparison of the Cochrane collaboration risk of bias tool and the effective public health practice project quality assessment tool: methodological research. J Eval Clin Pract 2012;18:12−8.

[15] Herbison P, Hay-Smith J, Gillespie WJ. Adjustment of meta-analyses on the basis of quality scores should be abandoned. J Clin Epidemiol 2006;59:1249−56.

[16] O'Connor SR, Tully MA, Ryan B, Bradley JM, Baxter GD, McDonough SM. Failure of a numerical quality assessment scale to identify potential risk of bias in a systematic review: a comparison study. BMC Res Notes 2015;8:224.

[17] Colle F, Rannou F, Revel M, Fermanian J, Poiraudeau S. Impact of quality scales on levels of evidence inferred from a systematic review of exercise therapy and low back pain. Arch Phys Med Rehabil 2002;83:1745−52.

[18] Jarde A, Losilla JM, Vives J. Suitability of three different tools for the assessment of methodological quality in ex post facto studies. Int J Clin Heal Psychol 2012;12:97−108.

[19] Whiting P, Rutjes AWS, Dinnes J, Reitsma JB, Bossuyt PMM, Kleijnen J. A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools. J Clin Epidemiol 2005;58:1−12.

[20] Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011;343:d5928.

[21] Jarde A, Losilla JM, Vives J, Rodrigo MF. Q-Coh: a tool to screen the methodological quality of cohort studies in systematic reviews and meta-analyses. Int J Clin Heal Psychol 2013;13:138−46.

[22] Sterne JA, Higgins JPT, Elbers RG, Reeves BC, The development group for ROBINS-I. Risk Of Bias In Non-randomized Studies of Interventions (ROBINS-I): detailed guidance. Available at: www.riskofbias.info. Accessed May 30, 2017.

[23] Whiting PF, Rutjes AWS, 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;155:529−36.

[24] Wells G, Shea B, O'Connell D, Peterson J. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses 2000. Available at: http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp. Accessed May 30, 2017.

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

[26] Hartling L, Milne A, Hamm MP, Vandermeer B, Ansari M, Tsertsvadze A, et al. Testing the Newcastle Ottawa Scale showed low reliability between individual reviewers. J Clin Epidemiol 2013;66:982−93.

[27] Lo CK-L, Mertz D, Loeb M. Newcastle-Ottawa Scale: comparing reviewers' to authors' assessments. BMC Med Res Methodol 2014;14:45.

[28] Oremus M, Oremus C, Hall GBC, McKinnon MC. Inter-rater and test−retest reliability of quality assessments by novice student raters using the Jadad and Newcastle−Ottawa Scales. BMJ Open 2012;2:e001368.

[29] Pan A, Sun Q, Okereke OI, Rexrode KM, Hu FB. Depression and risk of stroke morbidity and mortality: a meta-analysis and systematic review. JAMA 2011;306:1241−9.

[30] Jike M, Itani O, Watanabe N, Buysse DJ, Kaneita Y. Long sleep duration and health outcomes: a systematic review, meta-analysis and meta-regression. Sleep Med Rev 2018;39:25−36.

[31] Porcelli B, Pozza A, Bizzaro N, Fagiolini A, Costantini MC, Terzuoli L, et al. Association between stressful life events and autoimmune diseases: a systematic review and meta-analysis of retrospective case-control studies. Autoimmun Rev 2016;15:325−34.

[32] Xue J, Chen W, Chen L, Gaudet L, Moher D, Walker M, et al. Significant discrepancies were found in pooled estimates of searching with Chinese indexes versus searching with English indexes. J Clin Epidemiol 2016;70:246−53.

[33] Zheng Y, Wu X, Lin X, Lin H. The prevalence of depression and depressive symptoms among eye disease patients: a systematic review and meta-analysis. Sci Rep 2017;7:1−9.

[34] Cohen JA. Coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37−46.

[35] Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971;76:378−82.

[36] Feinstein AR, Cicchetti DV. High agreement but low Kappa: I. The problems of two paradoxes. J Clin Epidemiol 1990;43:543−9.

[37] Lantz CA, Nebenzahl E. Behavior and interpretation of the κ statistic: resolution of the two paradoxes. J Clin Epidemiol 1996;49:431−4.

[38] Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol 1990;43:551−8.

[39] Uebersax J. Statistical methods for diagnostic agreement 2015. Available at: http://www.john-uebersax.com/stat/agree.htm. Accessed October 25, 2017.

[40] Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol 1993;46:423−9.

[41] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159.

[42] Gamer M, Lemon J, Fellows I, Sing P. irr: various coefficients of interrater reliability and agreement. R package version 0.84; 2012. Available at: https://cran.r-project.org/package=irr.

[43] Kendall MG. A new measure of rank correlation. Biometrika 1938;30:81.

[44] Hartung J, Knapp G, Sinha BK. Statistical meta-analysis with applications. Hoboken, NJ, USA: John Wiley & Sons, Inc; 2008.

[45] Verhagen AP, de Vet HCW, Vermeer F, Widdershoven JWMG, de Bie RA, Kessels AGH, et al. The influence of methodologic quality on the conclusion of a landmark meta-analysis on thrombolytic therapy. Int J Technol Assess Health Care 2002;18:11−23.

[46] Balk EM, Bonis PAL, Moskowitz H, Schmid CH, Ioannidis JPA, Wang C, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. JAMA 2002;287:2973−82.

[47] Hartling L, Hamm M, Milne A, Vandermeer B, Santaguida PL, Ansari M, et al. Validity and inter-rater reliability testing of quality assessment instruments. Rockville, MD: Agency for Healthcare Research and Quality (US); 2012. Available at: https://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0077153/.

[48] Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 1998;352:609−13.

[49] Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995;273:408−12.

[50] Egger M, Juni P, Bartlett C, Holenstein F, Sterne J. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. Health Technol Assess 2003;7:1−76.

[51] Ahn S, Becker BJ. Incorporating quality scores in meta-analysis. J Educ Behav Stat 2011;36:555−85.

[52] Conn VS, Rantz MJ. Focus on research methods: research methods: managing primary study quality in meta-analyses. Res Nurs Health 2003;26:322−33.

[53] Morgan R, Sterne J, Higgins J, Thayer K, Schunemann H, Rooney ATK. A new instrument to assess risk of bias in non-randomised studies of exposures (ROBINS-E): application to studies of environmental exposure. Abstr. Glob. Evid. Summit. Cape Town: Cochrane Database of Systematic Reviews; 2017. https://doi.org/10.1002/14651858.CD201702.

[54] Turner RM, Spiegelhalter DJ, Smith GCS, Thompson SG. Bias modelling in evidence synthesis. J R Stat Soc Ser A Stat Soc 2009;172:21−47.

[55] Hernán MA. With great data comes great responsibility. Epidemiology 2011;22:290−1.

[56] Faggion CM. The rationale for rating risk of bias should be fully reported. J Clin Epidemiol 2016;76:238.

[57] Bhindi B, Wallis CJD, Boorjian SA, Thompson RH, Farrell A, Kim SP, et al. The role of lymph node dissection in the management of renal cell carcinoma: a systematic review and meta-analysis. BJU Int 2018;121:684−98.