Journal of Clinical Epidemiology 112 (2019) 28−35

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Risk of bias in nonrandomized studies of interventions showed low inter-rater reliability and challenges in its application

Silvia Minozzi[a,b,*], Michela Cinquini[c], Silvia Gianola[d], Greta Castellini[b,d], Chiara Gerardi[c], Rita Banzi[c]

[a]*Department of Epidemiology, Lazio Regional Health Service, via Cristoforo Colombo 112, 00147 Rome, Italy*
[b]*Department of Biomedical Sciences for Health, University of Milan, via Carlo Pascal 36, 20133 Milan, Italy*
[c]*Mario Negri Institute for Pharmacological Research IRCCS, Via Giuseppe La Masa 19, 20156, Milan, Italy*
[d]*IRCCS Istituto Ortopedico Galeazzi, Unit of Clinical Epidemiology, Via R.Galeazzi 4, 20162, Milan, Italy*

Accepted 8 April 2019; Published online 11 April 2019

Abstract

Objective: To assess the inter-rater reliability (IRR) and usability of the risk of bias in nonrandomized studies of interventions tool (ROBINS-I).

Study Design and Setting: We designed a cross-sectional study. Five raters independently applied ROBINS-I to the nonrandomized cohort studies in three systematic reviews on vaccines, opiate abuse, and rehabilitation. We calculated Fleiss' Kappa for multiple raters as a measure of IRR and discussed the application of ROBINS-I to identify difficulties and possible reasons for disagreement.

Results: Thirty one studies were included (195 evaluations). IRRs were slight for overall judgment (IRR 0.06, 95% CI 0.001 to 0.12) and individual domains (from 0.04, 95% CI −0.04 to 0.12 for the domain "selection of reported results" to 0.18, 95% CI 0.10 to 0.26 for the domain "deviation from intended interventions"). Mean time to apply the tool was 27.8 minutes (SD 12.6) per study. The main difficulties were due to poor reporting of primary studies, misunderstanding of the question, translation of questions into a final judgment, and incomplete guidance.

Conclusion: We found ROBINS-I difficult and demanding, even for raters with substantial expertise in systematic reviews. Calibration exercises and intensive training before its application are needed to improve reliability.

Keywords: Nonrandomized studies; ROBINS-I; Inter-rater reliability; Risk of bias; Systematic reviews

1. Introduction

Randomized controlled trials (RCTs) are considered the gold standard for the assessment of health interventions. However, in some areas of health care, RCTs may not be feasible for practical or ethical reasons or because relevant outcomes (e.g., long-term outcomes or rare events) may be hard to evaluate. Nonrandomized studies (NRS) can integrate the information from RCTs, by assessing the effectiveness of interventions on wider populations and with longer follow-up. Evidence from NRS is often put forward in evaluating the effectiveness of treatments. For instance, regulatory authorities recognize "real-world" observational studies as a necessary and important source of evidence in their decision-making [1−3]. About half of published systematic reviews (SRs) included NRS of intervention effects [4], and a very large number of guidelines relies on evidence from NRS. Thus, assessment of their validity is essential to enable clinicians and policy-makers to take decisions based on full evaluation of the strengths and weaknesses of the evidence.

Many tools have been proposed to assess the methodological quality of NRS in an SR. An overview published in 2007 found 87 different instruments to appraise methodological quality of NRS [5]. The Newcastle-Ottawa [6] and Downs-Black [7] are two of the most popular checklists addressing the methodological quality of study conduct and including also items relating to external validity.

In 2016, the risk of bias in nonrandomized studies of interventions (ROBINS-I) was launched to assess the risk of

Funding: This study did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of interest: None of the authors has competing interest to declare.

* Corresponding author. Cochrane Review Group on Drugs and Alcohol, Department of Epidemiology, Lazio Regional Health Service, Via Cristoforo Colombo 112, 00147 Rome, Italy. Tel.: +393286772798.

E-mail address: minozzi.silvia@gmail.com (S. Minozzi).

https://doi.org/10.1016/j.jclinepi.2019.04.001
0895-4356/© 2019 Elsevier Inc. All rights reserved.

What is new?

Key findings

- Applied to a sample of 31 studies by five raters, the IRRs for the overall judgment and for single domains were slight. The main difficulties were due to poor reporting of primary studies, misunderstanding of the question, translation of questions into a final judgment, and incomplete guidance.
- ROBINS-I was complex and demanding. Authors who use the tool must have good subject matter knowledge and be highly experienced epidemiologists in the conduction of nonrandomized studies.

What this adds to what was known

- Very few validation studies of ROBINS-I have been done as yet, so the assessment of the reliability, usability and experience can inform systematic reviewers and other users of the tool.

What is the implication and what should change now?

- To improve the proper application of the tool, users should be advised to run a calibration exercise and should be assisted by graphical algorithms, possibly implemented in software package. More examples taken from different settings could be provided in the guidance.

bias of comparative NRS of interventions [8]. It is applicable to cohort-like designs, such as observational cohort studies, quasirandomized trials, and other concurrently controlled studies. The tool is also applicable to case-control and cross-sectional studies, interrupted time series, and controlled before-after studies, although the authors are currently considering whether modifications of some signaling questions (SQs) are required for these designs.

The tool comprises seven domains and an overall judgment of risk of bias. Risk of bias for each domain and the overall judgment can be expressed as “low”, “moderate”, “serious”, critical”, or “no information” [9]. Each domain has several SQs to facilitate judgments about risk of bias. The response options to SQs are ”Yes”, “Probably yes”, Probably no”, “no” and “No information.” The risk of bias should be assessed for each outcome relevant to the SR.

Before applying the ROBINS-I to the studies in the SR, authors should define an RCT that would ideally answer the review question (target trial). They should also state whether the effect of interest is that of the assignment to the intervention(s), regardless of the extent to which they were actually received during the follow-up, or the effect of starting and adhering to the interventions as specified in the protocol. Finally, they should draft a list of critical confounding factors and cointerventions. To ensure proper application of the tool, both subject matter experts and method experts should be included in the review team.

As of January 2019, a rapid search on PubMed and Cochrane Library with the free text “ROBINS-I” yielded 79 non-Cochrane protocols or full SRs, 40 Cochrane protocols, and 26 Cochrane reviews that adopted the tool. Another 35 Cochrane and non-Cochrane reviews or protocols mentioned the use of A Cochrane Risk Of Bias Assessment Tool: for Non-Randomized Studies of Interventions, the first version of the tool, developed in 2014 [10]. The use of ROBINS-I is expected to increase given its official endorsement by Cochrane and GRADE [11].

We designed a cross-sectional study to evaluate the inter-rater reliability of ROBINS-I, the time needed to apply the tool, and the difficulties in its application that might help explain discrepancies.

2. Materials and methods

2.1. Sample size and selection

We estimated that a sample of 39 “evaluations,” that is, the number of evaluated outcomes per rater per each study, would be enough to obtain a 0.40 agreement with a reasonable margin of error of 0.4 [12,13]. This level of agreement was deemed acceptable, given the limited expertise of the raters with the tool, the complexity of NRS, and the variability of the topics.

We assessed the risk of bias of cohort-type NRS included in three reviews in the field of vaccines [14], opiate substitution for drug injectors [15], and ankle rehabilitation [16] (Table 1). These reviews did not use the ROBINS-I and enabled us to obtain a sample of NRS assessing the effect of the assignment and of starting and adhering to interventions. Finally, they reported results for subjective and objective outcomes. We based our assessment only on the primary publication cited in the reviews, as NRS protocols are not frequently registered or published.

2.2. Risk of bias assessment with ROBINS-I

Five raters were involved. All had extensive expertise in the conduct of SRs of RCTs, including the assessment of risk of bias, but medium or low expertise in the analysis and critical appraisal of NRS. Two of them had already used the ROBINS-I tool once, the others were naive. Only one had attended a workshop on the use of ROBINS-I. Two raters had a proof content knowledge in rehabilitation, one in the addiction field, and two in the vaccine field.

For each of the included review, raters defined if the objective was to assess the effect of the assignment or of

Table 1. Reviews and primary studies

Author, year	Topic	Effect assessed	No. of studies	No. of evaluations	Outcomes considered
Jefferson 2012 [14]	Efficacy of influenza vaccines on children	Assignment	14	90	Objective: influenza (laboratory confirmed) Subjective: Influenza-like syndrome
Mac Arthur 2012 [15]	Impact of opiate substitution treatment in people who inject drugs on HIV seroconversion	Starting and adhering	10	50	Objective: incidence of HIV seroconversion
Smeeing 2015 [16]	Efficacy of weight-bearing and mobilization in the postoperative care of ankle fractures	Starting and adhering	7	55	Objective: return to work Subjective: functional ability

starting and adhering to the intervention, the target RCT, and agreed on the critical confounding and cointerventions; then they independently read the full report of each study and applied the tool. They followed the ROBINS-I guidance [9] integrated by the graphic illustration of the SQ algorithms prepared by the authors (Appendix A); they did not do any calibration exercise or preliminary discussion on how to apply or interpret the tool. Raters assessed studies on vaccines first, then they were free to choose the order they preferred to assess the other studies.

Raters recorded the time needed to complete the tool in minutes, summing the time spent to read the full text and the time to complete the tool for each outcome.

After the calculation of the IRRs, the raters systematically discussed each domain and SQ in telephone meetings, to explore the possible reasons for lack of agreement. The outcome of these meetings was summarized and described narratively.

2.3. Data analysis

The number of evaluations was our unit of analysis. To measure the IRR, we calculated the Fleiss' Kappa for multiple raters for individual domains and for SQs. Agreement was classified as suggested by Landis and Koch [17]: values less than 0 indicate no agreement, 0 to 0.20 slight, 0.21 to 0.40 fair, 0.41 to 0.60 moderate, 0.61 to 0.80 substantial, and 0.81 to 1 almost perfect agreement. We assembled the ''Yes'' and "Probably Yes" answers and the "No" and Probably No'' answers'', as recommended in the ROBINS-I tool guidance [9].

We planned three subgroup analyses to explore reasons for disagreement by comparing the IRR of studies (1) assessing the effect of assignment only with those assessing the effect of starting and adhering, (2) measuring objective outcomes with those assessing subjective outcomes, and 3) covering different topics.

We used SAS 9.4 (SAS Institute, Cary, NC) to analyze the data. The data sets and scripts underlying the findings reported in this article are stored in Mendeley Data.

3. Results

3.1. Study characteristics

The five raters independently applied the ROBINS-I to 31 studies, for a total of 195 evaluations (Table 1). Appendix B reports the main characteristics of the studies and their references.

3.2. Inter-rater reliability

Figure 1 shows the IRRs for the overall final judgment of risk of bias and for single domains. The agreement for overall judgment was slight (IRR 0.06, 95% CI 0.001 to 0.12; 31 studies, 195 evaluations) as well as for individual domains, ranging from 0.04 (95% CI −0.04 to 0.12) for domain seven (selection of the reported results) to 0.18 (95% CI 0.10 to 0.26) for domain four (deviation from intended interventions). Agreement for each SQ ranged from no agreement (IRR −0.09, 95% CI −0.17 to −0.01) to almost perfect agreement (IRR 0.90 95% CI 0.82 to 0.98) (Appendix C).

As shown in Figure 2, agreement was slight in the overall judgment for subjective outcomes (17 studies, 85 evaluations), and there was no agreement for objective outcomes (22 studies, 110 evaluations). The agreement was slight for studies assessing the effect of assignment (14 studies, 90 evaluations), and there was no agreement for studies assessing the effect of starting and adhering (17 studies, 105 evaluations). Finally, comparing studies on different topics, we found slight agreement for studies on vaccines (14 studies, 70 evaluations) and rehabilitation (7 studies, 55 evaluations) and no agreement for studies on opiate substitution (10 studies, 50 evaluations).

3.3. Time needed to apply the tool

The mean time taken to apply ROBINS-I was 27.8 (SD 12.6) minutes per study.

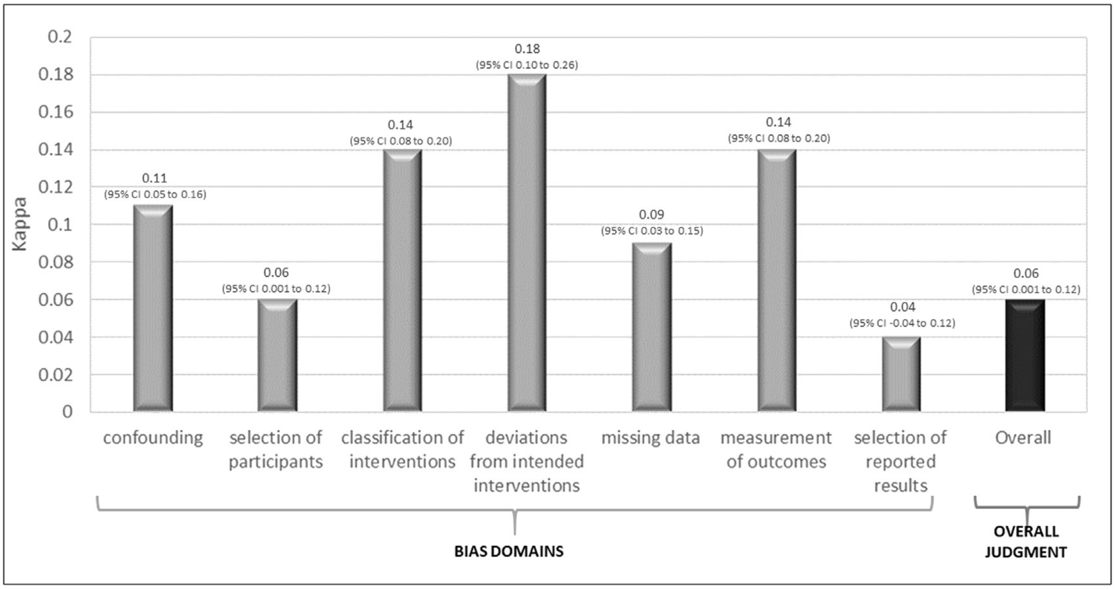

Fig. 1. Inter-rater reliability for overall judgment and single domains.

3.4. Reasons for poor agreement

Table 2 reports the main reasons for disagreement in each domain that emerged from the discussion among the raters. In summary, the poor agreement was linked to (1) misunderstanding of the question, possibly due to the limited expertise in the critical appraisal of NRS; (2) translation of the question in a final judgment, possibly due to the lack of experience with the tool; (3) poor quality of reporting of the assessed NRS; and (4) insufficient instructions in the guidance on decision trees of conditional SQs when the answer to the previous SQs is "no information."

The main issues with the interpretation of the questions were on domains one (bias due to confounding) and four (bias due to deviations from intended interventions). For instance, although clearly stated in the guidance, some raters had difficulty comparing the NRS to the target RCT and tended to express their judgment using a well-designed observational study as reference.

The concept of time-varying confounding applied to the NRS assessed in this analysis was also an issue. Similarly, the term "usual practice" and its interpretation (SQ 4.1) in the context of the assessed studies caused significant disagreement. In some cases (SQs 1.4, 1.5, 2.4, and 6.1),

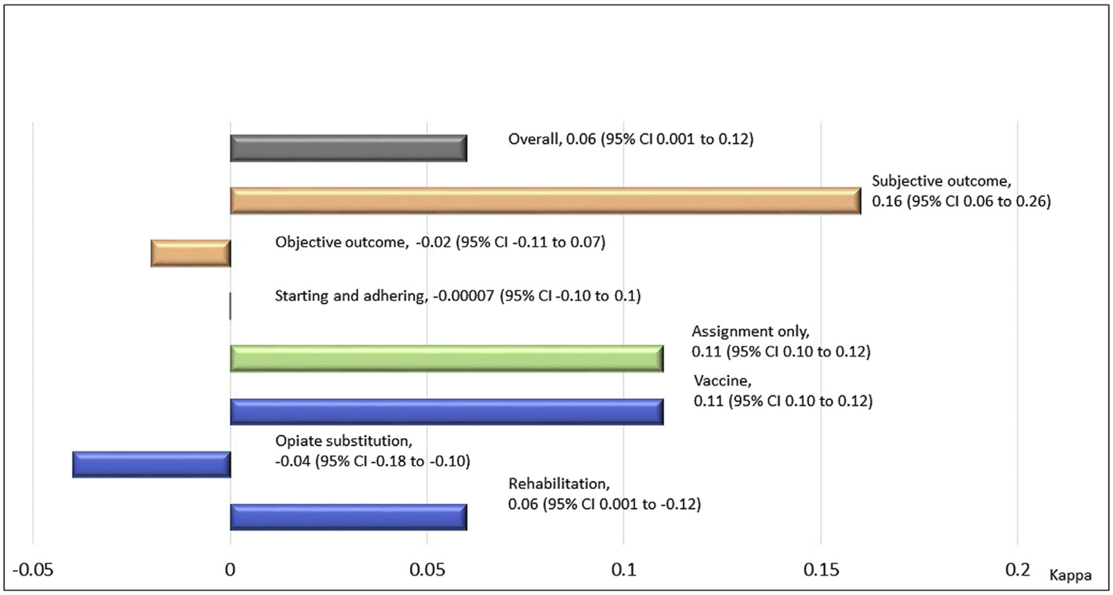

Fig. 2. Inter-rater reliability for overall judgment in subgroup analyses.

Table 2. ROBINS-I domains and signaling questions [9]: explanations for poor agreement found during discussion

Domain 1: Bias due to confounding	
1.1 Potential for confounding of the effect of intervention	Misunderstanding the question: some raters found difficulties comparing the nonrandomized studies to the target randomized controlled trials
1.2. Analysis based on splitting participants' follow-up time according to intervention received	Misunderstanding the question: not clear whether the question dealt with the risk of time-varying confounding or the analysis was actually based on splitting follow-up time.
1.3. Intervention discontinuations or switches likely to be related to factors that are prognostic for the outcome	Application of conditional questions: not clear how to apply this question on studies that assess assignment only to an intervention.
1.4. Appropriate analysis method that controlled for all the important confounding domains	Translation of question into a final judgment: differences on how strict to be in the judgment; some raters said ''no'' if only some confounders were included in the adjusted analysis, whereas others answered ''yes''
1.5. Confounding domains that were controlled for measured validly and reliably	Translation of question into a final judgment: differences on how to express the judgment when some confounders were measured validly and reliably and others were not
1.6. Control for any postintervention variables that could have been affected by the intervention	Poor quality of reporting of primary studies: not clear whether raters were merely asked to respond on the basis of the information provided, saying ''no information'', or if they should try to draw some inference or make guesses and say ''probably yes'' or ''probably no''
1.7. Appropriate analysis method that controlled for all the important confounding domains and for time-varying confounding	Low agreement on previous conditional questions: poor agreement at SQ 1.3
1.8. confounding domains that were controlled for measured validly and reliably	Low agreement on previous conditional questions: poor agreement at SQ 1.3
Domain 2: Bias in selection of participants into the study	
2.1. Selection of participants into the study (or into the analysis) based on participant characteristics observed after the start of intervention	Poor quality of reporting of primary studies
2.2. Postintervention variables that influenced selection likely to be associated with intervention	Poor quality of reporting of primary studies
2.3 Postintervention variables that influenced selection likely to be influenced by the outcome or a cause of the outcome	Poor quality of reporting of primary studies
2.4. Start of follow-up and start of intervention coincide for most participants	Translation of question into a final judgment: no agreement on the size of the time window between start of intervention and start of follow-up to give an answer of ''no''
2.5. Adjustment techniques used likely to correct for the presence of selection biases	
Domain 3: Bias in classification of interventions	
3.1 Intervention groups clearly defined	Poor quality of reporting of primary studies
3.2 Information used to define intervention groups recorded at the start of the intervention	Poor quality of reporting of primary studies
3.3 Classification of intervention status affected by knowledge of the outcome or risk of the outcome	Poor quality of reporting of primary studies
Domain 4: Bias due to deviations from intended interventions	
4.1. Deviations from the intended intervention beyond what would be expected in usual practice	Misunderstanding the question: not clear how to interpret the term ''usual practice'' and the amount of deviation, mainly for studies that assessed the effect of assignment only
4.2. Deviations from intended intervention unbalanced between groups and likely to have affected the outcome	Low agreement on previous conditional questions
4.3. Important cointerventions balanced across intervention groups	Poor quality of reporting of primary studies
4.4. Intervention implemented successfully for most participants	Misunderstanding the question: not clear of the difference between SQ 4.4 and 4.5
4.5. Participants adhere to the assigned intervention regimen	Misunderstanding the question: not clear of the difference between SQ 4.4 and 4.5
4.6. Appropriate analysis to estimate the effect of starting and adhering	
Domain 5: Bias due to missing data	
5.1 Outcome data available for all, or nearly all, participants	Poor quality of reporting of primary studies
5.2 Participants excluded because of missing data on intervention status	Poor quality of reporting of primary studies

(*Continued*)

Table 2. Continued

5.3 Participants excluded because of missing data on other variables needed for the analysis	Poor quality of reporting of primary studies
5.4 Proportion of participants and reasons for missing data similar across interventions	Poor quality of reporting of primary studies
5.5 Results were robust to the presence of missing data	Poor quality of reporting of primary studies
Domain 6: Bias in measurement of outcomes	
6.1 Outcome measure influenced by knowledge of the intervention received	Translation of question into a final judgment: disagreement on the weight given by raters to the methods for measuring objective outcomes
6.2 Outcome assessors aware of the intervention received by participants	Poor quality of reporting of primary studies
6.3 Methods of outcome assessment comparable across intervention groups	Poor quality of reporting of primary studies
6.4 Systematic errors in measurement of the outcome related to intervention received	Misunderstanding the question: the guidance does not provide any examples
Domain 7: Bias in selection of the reported results	
7.1. Reported effect estimate selected, on the basis of the results, from multiple outcome measurements within the outcome domain	Poor quality of reporting of primary studies
7.2. From multiple analyses of the intervention-outcome relationship	Poor quality of reporting of primary studies
7.3. From different subgroups	Poor quality of reporting of primary studies

the meaning was clear, but disagreement arose on how to translate the assessment into the final judgment. When only some of the critical confounders were measured validly and reliably, some raters were stricter than others (SQ 1.4 and 1.5). Similarly, raters differed in judging whether the follow-up and start of intervention coincided as they assessed minor differences in the time window differently (SQ 2.4).

The lack of adequate reporting affected agreement in almost all the domains, especially two, three, five and seven. Some raters attempted a judgment, whereas others responded strictly, based on the information provided, using the “no information” option.

Finally, as several SQs had to be answered depending on the answer to the previous one (conditional questions), the lack of agreement in some SQs caused divergent decision trees that contributed to the overall low reliability. The lack of clear instructions on which questions were applicable if the answer to the trigger question was “no information” caused major differences.

4. Discussion

4.1. Main findings

Five raters applied ROBINS-I to a sample of 31 cohort studies included in three SRs on different topics. The IRRs for overall judgment and for each domain were slight. Application of the ROBINS-I was time-consuming, requiring approximately half an hour for each study.

These results prompted several reflections. First, the difficulties in applying the tool and interpreting the questions can be ascribed to the lack of specific expertise in the analysis and critical appraisal of NRS. Raters had quite substantial experience in the assessment of risk of bias of RCTs, but this may not be sufficient, given the complexity and heterogeneity of NRS designs and analysis. We did not do any preliminary calibration exercises and relied only on the information in the ROBINS-I guidance [9]. Our analysis suggests that a thorough preliminary discussion on how to apply the tool, including a pilot run of the evaluation, would be useful. As clearly stated by the ROBINS-I developers, the range of expertise needed to apply the tool correctly should include methodology, statistics, and clinical aspects. It also calls for training on risk of bias of NRS and on the tool itself.

The tool’s complexity was also a major driver of low IRR. Raters found the pathway of conditional SQs complex and not always intuitive, making them hard to remember. We prepared graphic algorithms showing the question pathways (Appendix A), as the current version of the ROBINS-I guidance does not include them, although another tool for the assessment of risk of bias in RCTs [18]. These algorithms and their implementation in a software package may help the ROBINS-I users and possibly reduce errors. They may also reduce disagreement deriving from different interpretations, or translation of the concept, which was clear in principle, into a final judgment. More examples, from different fields of medicine, could be integrated into the ROBINS-I guidance to clarify the questions and highlight those situations where actual bias may arise. As presented in the ROBINS for the assessment of risk of bias in SRs [19], practical examples of the application on different studies could be useful too.

The lack of adequate reporting in most of the studies assessed in our analysis contributed to the poor agreement.

The impact of poor-quality reporting on the risk of bias assessment is common to any study design. Although reporting guidelines for observational studies exist, like the STROBE and its extensions, their application by authors and endorsement by journals is still suboptimal [20]. Several questions of the ROBINS-I are very detailed, and the poor reporting actually prevented a firm assessment of most of the relevant bias. Some raters did attempt a judgment based on deductive reasoning, whereas others answered "no information" to several SQs. This issue could be discussed and a common approach agreed before applying the tool, but further clarification in the guidance would also be useful.

4.2. Our findings in the context

Our results differ from those of two similar studies in terms of IRRs, but the conclusions are substantially similar. In the study by Losilla et al. [21], two raters applied the tool on 28 studies on health psychology. They found an IRR ranging from slight to perfect for SQs (0.00 to 1) and single domains (0.08 to 0.93) and a moderate IRR (0.57) for the overall final judgment. However, the raters judged ROBINS-I as very poor for clarity of instructions and clarity of the items, very demanding in terms of the amount of information to be collected, when the reporting of NRS is often very poor and not up to the standards sets by these demands.

In the study by Bilandic et al. [22], two raters applied the first version of the tool (A Cochrane Risk Of Bias Assessment Tool: for Non-Randomized Studies of Interventions) on 39 studies taken from two reviews on internal medicine. They found an IRR ranging from moderate to perfect for single domains (0.50 to 1.00) and from substantial to perfect for overall judgment (0.72 to 0.91). However, the time to complete the tool ranged from 4 to 2.5 hours per study. The raters judged the tool demanding to use and concluded "proper application of the instrument requires a substantial time and resource commitment in addition to an in-depth understanding of the sources of bias in NRS".

Differences in the IRRs between these studies and our study probably arose because they did a calibration exercise on a couple of studies before applying the tool and all the raters had good knowledge of the subject matter of the studies assessed.

Finally, a recent application of the ROBINS-E tool [23], the adaptation of the tool to exposure studies, found several limitations both in the theoretical construct and in the practical applicability [24].

4.3. Study limitations

One of the limits of our study is linked to the statistical analysis. We used Fleiss' Kappa, an extension of Scott's π for multiple raters and nominal variables. This statistic measures the difference between what is expected by chance and what has been observed. The expected agreement can exceed the observed agreement and then generate kappa values lower than 0. This is why, in some cases, Fleiss' Kappa may return low values even with actual high agreement (Fleiss' Kappa paradox) [25]. Our results might be further affected by this paradox.

4.4. Concluding remarks

Whether or not to include NRS in SRs remains an open question. Often concerns about potential bias inherent to nonexperimental studies are among the reasons for excluding them [26]. The definition of a specific framework for integrating evidence from NRSs with that from RCTs is one of the priorities of guideline developers. Schunemann et al. suggested a framework in which authors can use NRS as a complement, sequence, or replacement for RCTs by focusing on judgments about the population, intervention, comparison and outcomes [27]. As many public health and health service questions would probably fit this framework, there is a pressing need for reliable and useable tools to assess the risk of bias of NRSs.

Although the ROBINS-I tool is a comprehensive and detailed instrument, its application is demanding and complex. It requires a thorough knowledge of the subject matter and of the methodology for the conduct of NRS, including statistical issues. Given the high level of methodological, statistical, and content expertise and the need of intensive training, probably few groups worldwide might have the skills and resources to apply the tool in an appropriate and reliable way. This could make its implementation difficult, costly and hard to realize, so that limiting the possibility that this tool could become the tool of choice for NRS. Actions to possibly improve the reliability and usability of the tool may be implemented by both reviewers and ROBINS developers. The former should be trained on the tool and NRS methodology and encouraged to plan a calibration exercise before applying the tool. The latter should clarify SQs pathways, especially when quality of reporting is poor and the answer to SQs is "no information" and possibly reduce the number of SQs. Algorithms, better if implemented in a software package, may be helpful, as well as more practical examples of biases and application of the tool to different types of studies and fields of medicine.

CRediT authorship contribution statement

Silvia Minozzi: Conceptualization, Data curation, Investigation, Methodology, Project administration, Writing - original draft. **Michela Cinquini:** Data curation. **Silvia Gianola:** Investigation, Writing - review and editing. **Greta Castellini:** Investigation, Writing - review and editing. **Chiara Gerardi:** Investigation, Writing - review and editing. **Rita Banzi:** Conceptualization, Investigation, Methodology, Project administration, Writing - original draft.

Acknowledgments

The authors thank Judith Baggott for editing the article.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2019.04.001.

References

[1] Higgins JPT, Altman DG, Sterne JAC. Assessing risk of bias in included studies. In: Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0. The Cochrane Collaboration; 2011. Available at www.cochrane-handbook.org. (Accessed January 31, 2019).

[2] Black N. Why we need observational studies to evaluate the effectiveness of health care. BMJ 1996;312:1215−8.

[3] Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-world evidence - what is it and what can it tell us? N Engl J Med 2016;375:2293−7.

[4] Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ 2017;358:j4008.

[5] Sanderson S, Tatt ID, Higgins JP. 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.

[6] Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses 2008. Available at http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp: (Accessed January 31 2019).

[7] Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. J Epidemiol Community Health 1998;52:377−84.

[8] 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-randomised studies of interventions. BMJ 2016;355:i4919.

[9] Sterne JAC, 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 http://www.riskofbias.info. (Accessed January 31 2019).

[10] Sterne JAC, Higgins JPT, Reeves BC. A Cochrane Risk Of Bias Assessment Tool: for Non-Randomized Studies of Interventions (ACROBAT-NRSI), Version 1.0.0 2014. Available at http://www.riskofbias.info: (Accessed January 31 2019).

[11] Schunemann HJ, Cuello C, Akl EA, Mustafa RA, Meerpohl JJ, Thayer K, et al. GRADE guidelines: 18. How ROBINS-I and other tools to assess risk of bias in nonrandomized studies should be used to rate the certainty of a body of evidence. J Clin Epidemiol 2018. [Epub ahead of print].

[12] Gwet KL. Variance estimation of nominal-scale inter-rater reliability with random selection of raters. Psychometrika 2008;73:407−30.

[13] Gwet KL. Handbook of Inter-Rater Reliability. 2nd Edition. USA: Advanced Analytics LLC; 2010.

[14] Jefferson T, Rivetti A, Di Pietrantonj C, Demicheli V, Ferroni E. Vaccines for preventing influenza in healthy children. Cochrane Database Syst Rev 2012;CD004879.

[15] MacArthur GJ, Minozzi S, Martin N, Vickerman P, Deren S, Bruneau J, et al. Opiate substitution treatment and HIV transmission in people who inject drugs: systematic review and meta-analysis. BMJ 2012;345:e5945.

[16] Smeeing DP, Houwert RM, Briet JP, Kelder JC, Segers MJ, Verleisdonk EJ, et al. Weight-bearing and mobilization in the postoperative care of ankle fractures: a systematic review and meta-analysis of randomized controlled trials and cohort studies. PLoS One 2015;10:e0118320.

[17] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159−74.

[18] Higgins JPT, Savović J, Page MJ, Sterne JAC, on behalf of The ROB2 Development Group. Revised Cochrane risk-of-bias tool for randomized trials (RoB 2). Available at www.riskofbias.info. (Accessed January 31 2019).

[19] Whiting P, Savovic J, Higgins JP, Caldwell DM, Reeves BC, Shea B, et al. ROBIS: a new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol 2016;69:225−34.

[20] Sharp MK, Tokalic R, Gomez G, Wager E, Altman DG, Hren D. A cross-sectional bibliometric study showed suboptimal journal endorsement rates of STROBE and its extensions. J Clin Epidemiol 2018;107:42−50.

[21] Losilla JM, Oliveras I, Marin-Garcia JA, Vives J. Three risk of bias tools lead to opposite conclusions in observational research synthesis. J Clin Epidemiol 2018;101:61−72.

[22] Bilandzic A, Fitzpatrick T, Rosella L, Henry D. Risk of bias in systematic reviews of non-randomized studies of adverse cardiovascular effects of thiazolidinediones and cyclooxygenase-2 inhibitors: application of a new Cochrane risk of bias tool. Plos Med 2016;13:e1001987.

[23] ROBINS-E tool (Risk Of Bias In Non-randomized Studies - of Exposures). Preliminary risk of bias for exposures tool template. Available at. http://www.bristol.ac.uk/population-health-sciences/centres/cresyda/barr/riskofbias/robins-e/: (Accessed January 31 2019).

[24] Bero L, Chartres N, Diong J, Fabbri A, Ghersi D, Lam J, et al. The risk of bias in observational studies of exposures (ROBINS-E) tool: concerns arising from application to observational studies of exposures. Syst Rev 2018;7:242.

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

[26] Seida J, Dryden DM, Hartling L. The value of including observational studies in systematic reviews was unclear: a descriptive study. J Clin Epidemiol 2014;67:1343−52.

[27] Schunemann HJ, Tugwell P, Reeves BC, Akl EA, Santesso N, Spencer FA, et al. Non-randomized studies as a source of complementary, sequential or replacement evidence for randomized controlled trials in systematic reviews on the effects of interventions. Res Synth Methods 2013;4:49−62.