Assessing the quality of studies in meta-research:
review/guidelines on the most important quality assessment tools

Claudio Luchini[1*], Nicola Veronese[2], Alessia Nottegar[3], Jae Il Shin[4], Giovanni Gentile[5], Umberto Granziol[6], Pinar Soysal[7], Ovidiu Alexinschi,[8] Lee Smith[9*], Marco Solmi[4]

[1] Department of Diagnostics and Public Health, University and Hospital Trust of Verona, 37134 Verona, Italy
[2] National Research Council, Neuroscience Institute, Aging Branch, 35131 Padova, Italy
[3] Department of Diagnostics, Section of Pathology, San Bortolo Hospital, 36100 Vicenza, Italy
[4] Department of Pediatrics, Yonsei University College of Medicine, Yonsei-ro 50, Seodaemun-gu, C.P.O. Box 8044, Seoul 03722, Republic of Korea
[5] Department of Neurosciences, University of Padova, 35122 Padova, Italy
[6] Department of General Psychology, University of Padova, 35122 Padova, Italy
[7] Department of Geriatric Medicine, Faculty of Medicine, Bezmialem Vakif University, Istanbul, Turkey
[8] Institute of Psychiatry "Socola", Iasi, Romania
[9] The Cambridge Centre for Sport and Exercise Sciences, Anglia Ruskin University, Cambridge, UK

*Co-corresponding Authors:

Prof. **Claudio Luchini**, MD, PhD
Department of Diagnostics and Public Health, Section of Pathology
University and Hospital Trust of Verona
Piazzale Scuro, 10, 37134 Verona, Italy
Phone: 0039.045.8124835
Fax: 0039.045.8127136
Email: claudio.luchini@univr.it

Dr. **Lee Smith**, PhD
The Cambridge Centre for Sport and Exercise Sciences
Anglia Ruskin University, Cambridge, UK

ABSTRACT

Systematic reviews and meta-analyses pool data from individual studies to generate a higher level of evidence to be evaluated by guidelines. These reviews ultimately guide clinicians and stake-holders in health-related decisions. However, the informativeness and quality of evidence synthesis inherently depends on the quality of what has been pooled into meta-research projects. Moreover, beyond the quality of included individual studies, only a methodologically correct process, in relation to systematic reviews and meta-analyses themselves, can produce a reliable and valid evidence synthesis. Hence, quality of meta-research projects also affect evidence synthesis reliability. In this overview, the authors provide a synthesis of advantages and disadvantages, and main characteristics of some of the most frequently used tools to assess quality of individual studies, systematic reviews and meta-analyses. Specifically, the tools considered in this work are the Newcastle-Ottawa Scale (NOS) and the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) for observational studies, the Consolidated Standards of Reporting Trials (CONSORT), the Jadad scale, the Cochrane risk of bias tool 2(RoB2) for randomized controlled trials, the Preferred Reporting Items for Systematic Reviews and Meta-analysis (PRISMA) and the Assessment of Multiple Systematic Reviews (AMSTAR2), and AMSTAR-PLUS for meta-analyses.

KEYWORDS

Quality; meta-research; meta-analysis; NOS; STROBE; CONSORT; Cochrane; AMSTAR2; AMSTAR-PLUS; PRISMA.

INTRODUCTION

The quality assessment of individual studies included in meta-research manuscripts represents a fundamental step for supporting all the evidence synthetized by meta-research [1,2]. Although meta-research is considered the highest scientific level for summarizing the results from different analyses, at the same time there are also many potential sources of biases [3-5]. These sources of bias include for instance a potentially imprecise selection of subjects, a potentially inaccurate data collection and analysis, and possible biases in reporting the results of studies. It should be noted that the majority of potential biases derive directly from the manuscripts that are selected for a given meta-research study, hence the assessment of their quality must be accurately evaluated, using validated and standardized tools [1,2].

Producing reliable data to support results of scientific research and medical and public health decisions is of growing importance, given the huge rise in scientific reports published each year across different areas of medicine, and the frequently conflicting results of studies on the same topic. To summarize the effectiveness of medical interventions for a disease, and evidence from different studies that is conflicting, the only criterion to rank sources of information is the quality of studies, which need to be assessed using specific validated instruments [1].

This overview aims to serve as a starting point and a brief guide to identify and understand the main and most frequently used tools for assessing the quality of studies included in meta-research. The authors here share their experience in publishing several meta-research related articles covering different areas of medical sciences, including pathology, oncology, psychiatry, internal medicine, geriatrics, sport-exercise medicine, cardiology and nutrition.

The tools that will be discussed in this review include: the Newcastle-Ottawa Scale (NOS) and the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE), for observational studies; The Consolidated Standards of Reporting Trials (CONSORT), the Jadad scale, and the Cochrane risk of bias tool, for randomized controlled trials; The Preferred Reporting Items for Systematic Reviews and Meta-analysis (PRISMA), together with the Assessment of Multiple Systematic Reviews 2 and –PLUS (AMSTAR2 and AMSTAR-PLUS) for meta-analyses. For each item, a brief description, together with key strengths and limitations, will be discussed. Moreover, we provide examples of appropriate use of the included tools across different disciplines.

OBSERVATIONAL STUDIES

An observational study draws inferences from a group of subjects where the independent variable is not under the control of the researchers because of ethical concerns, operational restrictions or other reasons. In our experience, there are two main tools for quality assessment of observational studies, namely the "Newcastle-Ottawa Scale" (NOS) and the "Strengthening the Reporting of Observational Studies in Epidemiology" (STROBE).

The "Newcastle-Ottawa Scale" (NOS)

A comprehensive manual and additional information on this scale can be freely downloaded at: http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp.

Definition

The NOS can be used to assess the quality of the observational studies included in a meta-research analysis [6-10]. This tool can be used for both case-control and longitudinal (prospective, cohort studies) studies [9]. The NOS contains three domains, which are associated with the quality concept. These domains include (i) the selection of the participants, (ii) the comparability between cases and controls, and (iii) the accuracy of the evaluation of the outcomes. These domains are divided into eight specific items, which slightly differ when scoring case control and longitudinal studies [5-10]. Each item on the NOS is typically scored one point, except for the comparability item, which can be scored two points, after a specific adaptation to topic of interest [10]. Thus, the maximum score for each study is represented by 9 points, with studies having less than 5 points being identified as high risk of bias. [9-10]

Since this process may be subjective (e.g. the choice of the items to be scored is guided above all by the expertise of authors), it has been suggested that two independent researchers should score each paper. In our opinion, the most important as well as critical point in the NOS scoring relates to the common instruments utilized in specific fields in which the meta-analysis has been conducted. For instance, while exposure in studies focusing on subjects with acute myocardial infarction have used valid and objective measures, this might not always be the case in studies focusing on subjective constructs such a stress, depressive symptoms rated with self-report questionnaires, or with retrospective assessment of past events such as maltreatment or abuse.

Advantages and limitations

NOS has several advantages. First, NOS can be completed in a short amount of time. Second, it is the authors' view that the scale shows great adaptability to the investigated topic, owing to the versatile nature of its indexes. For example, different meta-analyses investigating the prognostic role of the same moderator (e.g. extranodal extension of nodal metastasis) in different cancer types can be evaluated adapting the NOS to the specific cancer (e.g. 1 point for a follow-up longer than 60 months for tumors with low or intermediate malignant behavior, such as thyroid, breast prostate or colorectal cancer [11-13], and 1 point for a follow-up longer than 36 months for more aggressive and highly malignant tumors, such as esophageal,

gastric or pancreatic cancer [11-16]). We believe that this practical model represents an excellent example of the NOS adaptability and versatility. Other not-negligible advantages are represented by the final score (ranging from 0 to 9), which can be used as a potential moderator in meta-regression analysis (or in sensitivity analyses), and its possible application for both case-control and longitudinal studies [9,10].

Limitations of the NOS are various and meta-researchers should be aware of these to limit their impact during the important process of quality assessment. First, some indices are not univocal, and some items need to be adapted when applying the NOS to cross-sectional and case-control studies. Second, although the adaptability of the indexes represents a point of strength, it can represent another possible source of bias. The points usually adapted by the authors are the number and type of adjustments in the multivariate analyses (if present), the duration of follow-up (as already discussed in the section of points of strength) and the outcome of interest not present at the baseline. Another potential limitation, in our opinion, is the suboptimal agreement that can be encountered between two independent reviewers in completing the NOS. Finally, it is not possible to apply NOS to cross-sectional studies that, however, are often included in meta-research.

Concluding remarks

The versatility of the NOS and its wide applicability are the most important advantages for using this quality assessment tool in meta-research. The correct interpretation and decision of the parameters to be listed and analyzed for the different indexes (e.g. 5 year of follow-up for prostate cancer and only 3 for the highly malignant pancreatic cancer), at least in our opinion, should be ensured through a complete knowledge of the topic analyzed within a meta-research, and the potential low agreement between coauthors in completing the NOS scale can be overcome through a final consensus, involving at least one additional expert coauthor.

The "Strengthening the Reporting of Observational Studies in Epidemiology" (STROBE)

A comprehensive manual and the downloadable form of STROBE checklist can be found at: https://www.strobe-statement.org/index.php?id=available-checklists.

Definition

STROBE is another tool that assists researchers in the fundamental step of quality assessment. It consists of a checklist of 22 items, which relate to the manuscript's title/abstract (1 item), introduction (2 items), methods (9 items), results (5 items), discussion (4 items) and findings (1 items) [17]. Eighteen items relate to cohort studies, case-control studies and cross-sectional studies, whereas four are specific to each of the three study designs. STROBE provides general reporting recommendations for descriptive observational studies and studies that investigate associations between exposures and health outcomes [17,18]. STROBE addresses the three main types of observational studies: cohort, case-control and cross-sectional studies. Recently, for improving scientific reporting in nutritional epidemiological studies, a new tool called STROBE-nut (STrengthening the Reporting of Observational Studies in Epidemiology-nutritional epidemiology) (STROBE-nut) has been designed [19]. This statement comprises a set of 24 items, organized as a checklist,

with the aim to ensure that all information is available, to enable quality appraisal, correct understanding, effective replication and application of findings. This tool can enhance the quality of the nutritional epidemiology field output. The additional items have been specifically designed for topics regarding nutrition-related issues and represent the direct demonstration of the adaptability of STROBE to different types of research areas.

Advantages and limitations

STROBE recommendations aim at explaining how to report research and provides detailed explanations for each checklist item [17-20]. STROBE may also aid in planning observational studies, and guide peer-reviewers and editors in their evaluation of manuscripts [17-20]

A key limitation of STROBE is that it does not specifically address topics such as genetic linkage studies, infectious disease modelling or case reports and case series. Thus, STROBE should be avoided for observational studies that specifically investigate diagnostic tests, tumor biomarkers and genetic associations. A recent paper has also highlighted that the endorsement of STROBE by journals is key to authors' awareness and adherence of the STROBE guideline [21]. Also, ambiguity in the language can affect STROBE reliability [22].

Concluding remarks

STROBE is not applicable to all possible observational studies. Therefore, authors carrying out meta-research must carefully consider its use to avoid incorrect choice of this item. At the same time, where applicable, it is a reliable and reproducible tool for assessing the quality of observational papers in meta-research. In our opinion, STROBE is a useful tool for double-checking if all the points required by an observational study are included in the manuscript. Unambiguous language is desirable to increase adherence in following guidelines and improve the quality of reporting.

RANDOMIZED CONTROLLED TRIALS

A randomized controlled trial (RCT) is a scientific experiment aiming at reducing potential sources of bias when testing the effectiveness of new treatments; this is accomplished by randomly allocating subjects to two or more groups, treating them differently, and then comparing the different treatments with respect to a measured response. Among others, three tools are frequently used in the quality assessment of randomized controlled trials in meta-research. They are: the "Consolidated Standards of Reporting Trials", the "Jadad scale", and the "Cochrane-risk of bias" tool 2 (RoB2).

The "Consolidated Standards of Reporting Trials" (CONSORT)

A comprehensive manual with additional information (to that discussed here) and the downloadable form of the CONSORT checklist can be found at http://www.consort-statement.org/.

Definition

CONSORT consists of a checklist of fundamental items that should be included in reports of randomized controlled trials and a diagram for documenting the flow of participants through a trial [23]. CONSORT can be adapted to a wider class of trial designs, such as equivalence, factorial, cluster, crossover and non-inferiority trials [23]. The main aim of the CONSORT is to furnish basic guidelines to authors for improving and enlightening the reports of RCTs. Indeed, they must be clear, complete and transparent, since all readers, peer reviewers and editors can also use CONSORT, to help them in critically appraising and interpreting reports of RCTs [23,24].

CONSORT has been extended recently to enhance the reporting of randomized adaptive-design clinical trials [24[. An Adaptive designs CONSORT Extension (so-called "ACE") guideline was developed, with the intention of enhancing transparency and improving the report of adaptive-design randomized trials, to increase both the interpretability of their results and reproducibility of their methods, results and inference [25].

Advantages and limitations

CONSORT represents an excellent instrument for ensuring, but also for clearly evaluating, the quality of RCTs.

The main limitation of CONSORT regards its initial design. Indeed, CONSORT was not designed n to be used as a quality assessment instrument [23-28]. Conversely, the content of CONSORT focuses on items specifically associated to the internal and external validity of trials. Several items not explicitly taken into account by CONSORT should also be cited in a report, such as information about ethical approval (including guidelines stated in the Declaration of Helsinki 1964 and subsequent amendments, and ethical committee approval), obtaining informed consent from participants, and, where relevant, existence of a data safety and monitoring committee [27]. In addition, any other aspects of a trial that are mentioned should be accurately described, as for example information regarding cost/effectiveness analysis.

Concluding remarks

CONSORT was not specifically conceived as an instrument for quality assessment; its content and application, however, is fundamental in meta-research papers investigating randomized controlled trials. This tool is continuously evolving in relation to covering new aspects of randomized controlled trials. Our recommendation is that authors of meta-research regularly review the literature around this topic to inform their effective implementation of the tool

The Jadad Scale

The seminal paper that introduced the Jadad scale was published in 1996 in Control Clin Trials [4].

Definition

Jadad scale is composed of three questions, with binary yes/no answers. The first question focuses on randomization, the second on double-blinding, the third on whether the study reported all information on non-completers. In addition to the three points given by a yes answer to the above three questions, two additional points can be obtained if authors described in detail a correct procedure for randomization, and for blinding. The maximum score of the Jadad scale is five.

Advantages and limitations

Based on our experience, the main advantage of Jadad scale is the short amount of time needed to complete it while still assessing main sources of bias. The main disadvantage of the tool is it neglects some key information on potential confounding factors affecting the validity of findings, such as allocation concealment, industry sponsorship, and conflict of interest. Moreover, the Jadad scale bases its score only on a few questions (which is also its advantage), without a clearer framework to be used as reference, and without more specific questions, and so can generate somewhat subjective ratings, at least compared to more comprehensive tools such a s RoB 2 (see below).

Concluding remarks

Jadad scale is a valid tool to initially rate the quality of an RCT, in a short amount of time. Jadad is a valid instrument for use when assessing RCTs. However, it predominantly provides a broad assessment. Therefore, other instruments have been developed to provide a more detailed assessment of RCTs. Indeed, its accuracy and precision are to date inferior compared with other more exhaustive updated instruments (see below).

The Cochrane Risk of Bias Tool 2 (RoB2)

The full manual and rationale behind the Cochrane Risk of Bias Tool 2 (RoB2) [29,30], together with video tutorials explaining how to score each item of the tool, as well as a Microsoft Excel sheet to score RoB2 can be found at: https://sites.google.com/site/riskofbiastool/welcome/rob-2-0-tool?authuser=0. Version 2 of this tool replaces the first version, originally published in 2008, and updated in 2011 [31].

Definition

The RoB2 tool assesses five domains, namely (1) risk of bias arising from randomization process, (2) risk of bias due to adherence or assignment to the intended intervention, (3) risk of bias due to missing outcome data, (4) risk of bias in measurement of outcome, (5) risk of bias in the selection of reported results. Each domain needs to be assessed by several questions: Three questions are asked for the *randomization domain* relating to (i) random sequence generation, (ii) allocation concealed until participants enrolled, and (iii) difference in baseline values. Seven questions are asked for the *adherence or assignment to intervention domain* relating to (i) participants blinding, (ii) person delivering treatment blinding, (iii) deviations from intended treatment (iv) type of analysis accounting for group assignment, and (v) potential errors in assigning an individual's results to the wrong group., Three questions are asked for the *bias due to missing outcome data domain* relating to (i) data available for almost all randomized subjects, (ii) result not biased by missing data, (iii) possibility or likeliness of missingness of outcome is related to the outcome's nature. Five questions are asked for the assessing *bias in outcome assessment domain* relating to (i) method of outcome measurement, (ii) difference in outcome assessment between groups, (iii) outcome assessor blinding, and (iv) blinding of participants.) Three questions are asked for the *assessing bias in results selective reporting domain* relating to (i)a-priori protocol, (ii) the presence of one or more measures to assess outcome, and (iii) the presence of one or more analyses to assess outcome. Each domain is rated as having a low risk of bias, as having some concerns suggesting bias, or as having a high risk of bias. Finally, an overall judgement of the risk of bias of a given RCT is provided, with a low risk of bias if low risk of bias is measured in all domains, some concerns if some concerns are measured in at least one domain, and high risk of bias if high risk of bias is measured in at least one domain.

Advantages and limitations

RoB2 form the Cochrane group is to our knowledge the most updated, reliable, valid, and comprehensive tool to assess potential biases in RCTs. Clear instructions are provided, together with instruments to be used to rate RoB2. Of course, such an exhaustive instrument takes significant time to rate one single RCT (which could be the only disadvantage compared with Jadad). However, we believe RoB2 largely outperforms Jadad methods, and should be considered as the gold-standard to rate RCTs quality.

Concluding remarks

RoB2 is the gold-standard instrument to evaluate the quality and presence of bias in RCTs. We recommend researchers take the necessary time to become confident with the instrument and given its complexity two blinded raters should score in parallel. RoB2 scored in double-blind should be the first choice for any high-quality meta-research project.

META-ANALYSES

Meta-analyses are meta-research projects which pool data from previously published individual studies, and provide a pooled estimate of the association investigated in original included studies. A meta-analysis accounts for within and between studies heterogeneity, and for random-error, when a random-effect model is chosen in analyses [32]. A meta-analysis also provides a heterogeneity index, which can be considered as a proxy of underlying factors which influence the effect size of individual studies. One of these factors that most frequently influence the magnitude of an effect size is the quality of studies included in a meta-analysis.

Preferred Reporting Items for Systematic Reviews and Meta-analysis (PRISMA)

A full manual and rationale to use PRISMA are available at http://www.prisma-statement.org/. PRISMA has been published in different journals, for instance in BMJ [33].

Definition

PRISMA works as a checklist to guide authors in correct reporting of all needed details for high quality reporting in systematic reviews and meta-analyses. PRISMA assesses the quality of reporting, not of the method used in the meta-analysis, neither on the quality of included studies. Similar, to the STROBE checklist, the PRISMA requests authors to report a title, a structured summary, rationale and objectives in the introduction. In the methods, authors must indicate at what page they provided information on protocol and its registration, eligibility criteria, information sources, search, study selection procedure, data collection process, data items when extracting data, risk of bias in individual studies, summary measures, synthesis of results, risk of bias across studies, and eventual additional analyses. In the results section authors must indicate study selection results (PRISMA flowchart), study characteristics, risk of bias within studies, results of individual studies, synthesis of results, risk of bias across studies, additional analyses. In the discussion, where they indicate summary of evidence, they must also indicate limitations and conclusions. Finally, PRISMA requests authors to disclose any sources of funding within the manuscript. For each of the above items of the checklist, authors must indicate at what page they reported the information.

Advantages and limitations.

Following the PRISMA statement it is a necessary step to provide high quality reporting in any systematic review or meta-analysis. We strongly suggest adhering to PRISMA statement for any meta-research process and in particular to use the PRISMA figure, which is the standardized optimal instrument to represent the flow-chart of study selection process.

Concluding remarks

PRISMA statement should be used as a guide when preparing a meta-research project, starting from its protocol. Also, it can be considered a measure of the quality in reporting of the meta-research project. However, it should not be used as an instrument for the quality assessment of a meta-analysis. Another point to be added, at least in our opinion, regards the use of different databases for meta-analysis, although it is not mandatory for PRISMA. For example,, it may be of importance to use different databases (e.g. Pubmed,

SCOPUS, Embase) to improve the overall quality and to expand the possibilities of finding all potentially useful references.

The Assessment of Multiple Systematic Reviews2 (AMSTAR2)

AMSTAR was first developed and launched in 2007. Starting from an initial 37 items, an exploratory factor analysis was used to identify underlying components producing 11 constructs that now make up the AMSTAR [34]. AMSTAR2 [35] is an updated and improved version of the AMSTAR [34,36,37], which is available at: https://amstar.ca/index.php , where it can be scored on-line. Of importance, AMSTAR2 has 16 items in total (compared to the 11 original), has simple categorization and indication for each domain, and has an overall rating scale (critically low to high), contrary to the original version.

Definition

AMSTAR2 is composed of 16 questions. The questions focus on whether authors considered PICO (Population Intervention Comparison Outcome) for inclusion criteria, if there was an "a-priori" protocol, if the reason to include one study design only was provided, if a comprehensive literature search was run (at least two databases searched), if study selection and data extraction were performed in duplicate, if a list of both included and excluded studies (after full-text assessment) was provided, if included studies are described in adequate detail, if authors assessed the risk of bias of included studies, if source of funding was stated, if statistical analyses were correct, if authors considered the impact of the risk of bias in the meta-analysis and when interpreting results, if authors explained heterogeneous results, if authors assessed and discussed publication bias, and finally if conflict of interest was disclosed. After answering the 16 questions of AMSTAR2 on the on-line platform, the quality (namely the confidence in results) is rated into either high, moderate, low, critically low.

Advantages and limitations

AMSTAR2 is a valid instrument to assess the methodological quality of meta-analyses and can be easily scored on a dedicated platform which also calculates the quality category; this instrument has been validated and is already widely used and deemed reliable by the scientific community. AMSTAR2 is not intended to generate an overall score. However, as also emerged by our experience, a disadvantage of the AMSTAR2 is that it does not really provide information beyond the methodological quality of the meta-analysis, and specifically on the studies included in the meta-analysis.

Concluding remarks

AMSTAR 2 is a validated tool that should be used to score the methodological quality of meta-analyses. However, to also gain some insight into the quality of studies included in the meta-analysis, and hence in the validity of a meta-analysis results, other instruments should be used.

The Assessment of Multiple Systematic Reviews-PLUS (AMSTAR-PLUS)

AMSTAR-PLUS has been proposed in a recent overview of meta-analysis to score the quality of both the methods of the meta-analysis, and its content [38].

Definition

AMSTAR-PLUS is composed of 16 items. The first 11 items are those of the AMSTAR. Authors have then supplemented AMSTAR with AMSTAR-PLUS items, which ask whether the majority of studies were double-blinded RCTs, if the total number of participants was sufficiently large, if the pooled effect size was confirmed in the largest individual study, if observed cases analyses were performed, if the outcome was heterogeneous, and if there was publication bias.

Advantages and limitations

Differently from AMSTAR2, to our knowledge AMSTAR-PLUS has not undergone rigorous validation and standardization. However, we believe that it adds information on the content of a meta-analysis, which is crucial information that cannot be neglected when measuring credibility of results from a meta-analysis. For instance, technically accurate meta-analyses can be performed, but they may include low quality studies, heterogeneous results, affected by publication bias, based on small studies, and so on. Hence, we believe that, AMSTAR-PLUS might be the first choice when assessing the quality of meta-analyses including RCTs.

Concluding remarks

AMSTAR-PLUS merges the insight on the methodological process of a meta-analysis with its first eleven questions, and in addition to this also provides an insight into the quality of studies included in a meta-analysis. AMSTAR-PLUS total score can be considered a s a single score, or it can split into the methodological scores (first eleven questions), and the "Content" score (last six questions).

CONCLUSIONS

In this review, we have reported and summarized the most commonly used tools for assessing quality in observational and intervention studies as well as in meta-analyses. Overall, all these tools have important advantages and disadvantages.

When performing an individual study or a meta-research project, STROBE, CONSORT, and PRISMA check-lists should always guide the protocol draft and the reporting in publications. In observational studies, a brief instrument such as the NOS scale merges accuracy and time needed to score the instrument, and should be considered as a first-option when assessing the quality of observational studies in meta-research. When assessing quality of RCTs, RoB2 should be considered as the first option. When assessing meta-analyses quality, AMSTAR2 should be used if one is only interested in methodological quality of a meta-analysis, but AMSTAR-PLUS should be used when one is also interested in the quality of the meta-analysis' content.

Another important aspect is the plasticity of these tools. Ideally, we would like to have only one tool for assessing all kinds of observational studies, and only one for assessing the quality/risk of bias of intervention studies. For example, we do not have any specific tool for open-label trials or non-controlled trials that are, on the contrary, of importance in certain disciplines such as psychiatry or geriatrics. Future research should aim to develop such tools and in the meantime researchers should utilize this overview to inform the most appropriate tool(s) to use for their meta-research projects.

HIGHLIGHTS

What is already known:

The informativeness and quality of evidence synthesis inherently depends on the quality of what has been pooled into meta-research projects. Beyond the quality of included individual studies, only a methodologically correct process, in relation to systematic reviews and meta-analyses themselves, can produce a reliable and valid evidence synthesis.

What is new:

In this overview, the authors provide a synthesis of advantages and disadvantages, and main characteristics of some of the most frequently used tools to assess quality of individual studies, systematic reviews and meta-analyses.

Potential Impact:

This overview serves as a starting point and a brief guide to identify and understand the main and most frequently used tools for assessing the quality of studies included in meta-research. The authors here share their experience in publishing several meta-research related articles covering different areas of medical sciences.

REFERENCES

[1] Dreier M.Quality Assessment in Meta-analysis. In: Doi SAR, Williams GM, editors. Methods of Clinical Epidemiology. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013: 213-228.

[2] Burns PB, Rohrich RJ, Chung KC. The levels of evidence and their role in evidence-based medicine. Plast Reconstr Surg 2011; 128: 305-310

[3] Higgins JP, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, Savovic J, Schulz KF, Weeks L, Sterne JA; Cochrane Bias Methods Group; Cochrane Statistical Methods Group. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011; 343: d5928.

[4] Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, McQuay HJ. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials 1996; 17: 1-12

[5] Murad MH, Montori VM, Ioannidis JP, Jaeschke R, Devereaux PJ, Prasad K, Neumann I, Carrasco-Labra A, Agoritsas T, Hatala R, Meade MO, Wyer P, Cook DJ, Guyatt G. How to read a systematic review and meta-analysis and apply the results to patient care: users' guides to the medical literature. JAMA. 2014 Jul;312(2):171-9

[6] Stang A. Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses. Eur J Epidemiol. 2010 Sep;25(9):603-5.

[7] Lo CK, Mertz D, Loeb M. Newcastle-Ottawa Scale: comparing reviewers' to authors' assessments. BMC Med Res Methodol. 2014 Apr 1;14:45.

[8] Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M, Tugwell P. The Newcastle-Ottawa Scale (NOS) for assessing the quality if nonrandomized studies in meta-analyses, 2012. Available from: URL: http://wwwohrica/programs/clinical_epidemiology/oxfordasp

[9] Luchini C, Stubbs B, Solmi M, Veronese N. Assessing the quality of studies in meta-analyses: Advantages and limitations of the Newcastle Ottawa Scale. World J Meta-Anal2017; 5(4): 80-84

[10] Cook DA, Reed DA. Appraising the quality of medical education research methods: the Medical Education Research Study Quality Instrument and the Newcastle-Ottawa Scale-Education. Acad Med. 2015 Aug;90(8):1067-76. doi: 10.1097/ACM.0000000000000786.

[11] Veronese N, Nottegar A, Pea A, Solmi M, Stubbs B, Capelli P, et al. Prognostic impact and implications of extra-capsular lymph node involvement in colorectal cancer: a systematic review with meta-analysis. Ann Oncol 2016; 27:42–8.

[12] Veronese N, Luchini C, Nottegar A, Kaneko T, Sergi G, Manzato E, Solmi M, Scarpa A. Prognostic impact of extra-nodal extension in thyroid cancer: A meta-analysis. J Surg Oncol. 2015 Dec;112(8):828-33.

[13] Nottegar A, Veronese N, Senthil M, Roumen RM, Stubbs B, Choi AH, Verheuvel NC, Solmi M, Pea A, Capelli P, Fassan M, Sergi G, Manzato E, Maruzzo M, Bagante F, Koç M, Eryilmaz MA, Bria E, Carbognin L, Bonetti F, Barbareschi M, Luchini C. Extra-nodal extension of sentinel lymph node metastasis is a marker of poor prognosis in breast cancer patients: A systematic review and an exploratory meta-analysis. Eur J Surg Oncol 2016; 42:919-25.

[14] Veronese N, Fassan M, Wood LD, Stubbs B, Solmi M, Capelli P, Pea A, Nottegar A, Sergi G, Manzato E, Carraro S, Maruzzo M, Cataldo I, Bagante F, Barbareschi M, Cheng L, Bencivenga M, de Manzoni G, Luchini C. Extranodal extension of nodal metastases is a poor prognostic indicator in gastric cancer: a systematic review and meta-analysis. J Gastrointest Surg 2016; 20:1692-8.

[15] Luchini C, Wood LD, Cheng L, Nottegar A, Stubbs B, Solmi M, Capelli P, Pea A, Sergi G, Manzato E, Fassan M, Bagante F, Bollschweiler E, Giacopuzzi S, Kaneko T, de Manzoni G, Barbareschi M, Scarpa A, Veronese N. Extranodal extension of lymph node metastasis is a marker of poor prognosis in oesophageal cancer: a systematic review with meta-analysis. J Clin Pathol 2016; [Epub ahead of print].

[16] Luchini C, Veronese N, Pea A, Sergi G, Manzato E, Nottegar A, et al. Extra-nodal extension in N1-adenocarcinoma of pancreas and papilla of Vater: a systematic review and meta-analysis of its prognostic significance. Eur J Gastroenterol Hepatol 2016; 28:205–9.

[17] Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the reporting of observational studies in epidemiology (STROBE): Explanation and elaboration. PLoS Med. 2007;4:e297.

[18] Cuschieri S. The STROBE guidelines. Saudi J Anaesth. 2019 Apr;13(Suppl 1):S31-S34.

[19] Lachat C, Hawwash D. The addition of STROBE-nut to the EJCN instructions to authors: some considerations and caveats. Eur J Clin Nutr. 2020;74(4):657-658. doi:10.1038/s41430-020-0581-z

[20] Adams AD, Benner RS, Riggs TW, Chescheir NC. Use of the STROBE Checklist to Evaluate the Reporting Quality of Observational Research in Obstetrics. Obstet Gynecol. 2018;132(2):507-512.

[21] Sharp MK, Bertizzolo L, Rius R, Wager E, Gómez G, Hren D. Using the STROBE statement: Survey findings emphasized the role of journals in enforcing reporting guidelines. J Clin Epidemiol. 2019 Aug 6. [Epub ahead of print].

[22] Sharp MK, Tokalić R, Gómez 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. 2019 Mar;107:42-50.

[23] Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. J Am Med Assoc 1996;276:637e9.

[24] Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. Lancet 2001;357:1191e4.

[25] Dimairo M, Pallmann P, Wason J, Todd S, Jaki T, Julious SA, et al. The Adaptive designs CONSORT Extension (ACE) statement: a checklist with explanation and elaboration guideline for reporting randomised trials that use an adaptive design. BMJ. 2020;369:m115.

[26] Hopewell S, Clarke M, Moher D, Wager E, Middleton P, Altman DG, et al.CONSORT for reporting randomized controlled trials in journal and confer-ence abstracts: explanation and elaboration.PLoS Med2008;5:e20.

[27] Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, Elbourne D, Egger M, Altman DG; CONSORT. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. Int J Surg. 2012;10(1):28-55.

[28] Palmas W. The CONSORT guidelines for noninferiority trials should be updated to go beyond the absolute risk difference. J Clin Epidemiol. 2017 Mar;83:6-7.

[29] Higgins JPT, Sterne JAC, Savović J, Page MJ, Hróbjartsson A, Boutron I, Reeves B, Eldridge S. A revised tool for assessing risk of bias in randomized trials In: Chandler J, McKenzie J, Boutron I, Welch V (editors). Cochrane Methods. *Cochrane Database of Systematic Reviews* 2016, Issue 10 (Suppl 1). dx.doi.org/10.1002/14651858.CD201601.

[30] Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, Cates CJ, Cheng H-Y, Corbett MS, Eldridge SM, Hernán MA, Hopewell S, Hróbjartsson A, Junqueira DR, Jüni P, Kirkham JJ, Lasserson T, Li T, McAleenan A, Reeves BC, Shepperd S, Shrier I, Stewart LA, Tilling K, White IR, Whiting PF, Higgins JPT. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ* (in press).

[31] Higgins JP, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, Savovic J, Schulz KF, Weeks L, Sterne JA; Cochrane Bias Methods Group; Cochrane Statistical Methods Group. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ. 2011 Oct 18;343:d5928. doi: 10.1136/bmj.d5928.

[32] DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986 Sep;7(3):177-88.

[33] Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). *Preferred Reporting Items for Systematic Reviews and Meta-Analyses:* The PRISMA Statement. BMJ 2009;339:b2535.

[34] Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, Porter AC, Tugwell P, Moher D, Bouter LM. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol. 2007 Feb 15; 7:10. PMID: 17302989.

[35] Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, Moher D, Tugwell P, Welch V, Kristjansson E, Henry DA. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ. 2017 Sep 21;358:j4008

[36] Shea BJ, Hamel C, Wells GA, Bouter LM, Kristjansson E, Grimshaw J, Henry DA, Boers M.AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. J Clin Epidemiol. 2009 Oct; 62(10):1013-20. PMID: 19230606.

[37] Shea BJ, Bouter LM, Peterson J, Boers M, Andersson N, Ortiz Z, Ramsay T, Bai A, Shukla VK, Grimshaw JM. External Validation of a Measurement Tool to Assess Systematic Reviews (AMSTAR). PLoS ONE. 2007; 2(12): e1350. PMCID: PMC2131785.

[38] Correll CU, Rubio JM, Inczedy-Farkas G, Birnbaum ML, Leucht S. Efficacy of 42 Pharmacologic Cotreatment Strategies Added to Antipsychotic Monotherapy in Schizophrenia: Systematic Overview and Quality Appraisal of the Meta-analytic Evidence. 2017 Jul 1;74(7):675-684. doi: 10.1001/jamapsychiatry.2017.0624.

Table 1. Characteristics of the quality assessment instruments considered in this overview.

Instrument	For which studies	Main aim of the instrument	Structure of instrument	Quality measure categories	Main advantage	Main disadvantage	Cut-off for high quality
NOS	Observational	Scale	8 questions for a maximum score of 9 – One star (*) for each item, except from comparability, which can be scored of 2 stars (**)	Continuous score. Range 0-9	Plasticity / Adaptability. Not time consuming.	Potential subjectivity in defining and scoring items	A score of at least 6 equals no high risk of bias
STROBE	Observational	Checklist	22 items	Not applicable	It's a guide for both protocol drafting, and results reporting.	Does not cover all possible observational studies (e.g. diagnostic tests, genetic associations)	Not applicable
CONSORT	Randomized controlled trials	Checklist	25 questions	Not applicable	It's a guide for both protocol drafting, and results reporting.	Some items not included, originally not a quality tool	Not applicable
RoB2	Randomized controlled trials	Assess the presence of bias	Five domains – several questions within each domain.	Low risk of bias Some concerns High risk of bias	Validated, reliable, comprehensive and fruible documentation to learn to use it.	Time consuming	See categories
JADAD	Randomized controlled trials	Assess the presence of bias	5 questions "yes"/"no"	Continuous score. Range 0-5	Valid for rough evaluation. Not time consuming.	Neglects some important sources of bias	$\geq$ 3/5
PRISMA	Meta-analyses	Guide and assess the quality of reporting	27 questions	Not applicable	It's a guide for both protocol drafting, and results reporting.	Some items not included, originally not a quality tool	Not applicable
AMSTAR2	Meta-analyses	Assess the methodological quality	16 items	High quality, moderate quality, low quality, critically low quality.	Validated instrument, online platform for scoring.	Does not provide information on the meta-analysis content.	See categories
AMSTAR-PLUS	Meta-analyses	Assesse both methodological quality, and content's quality	17 items	Continuous score 0 to 20. Two subscores for AMSTAR (1 to 11), and PLUS-Content (o to 9)	Assess both methodological quality and quality of included studies. Continuous score has been used as moderator in meta-regression.	Not validated.	None proposed to date.