Journal of Clinical Epidemiology 190 (2026) 112056

METHODOLOGICAL ASPECTS OF RESEARCH INTEGRITY AND CULTURE

A scoping review of critical appraisal tools and user guides for systematic reviews with network meta-analysis: methodological gaps and directions for tool development

K.M. Mondragon[a,*], C.S. Tan-Lim[a], R. Velasco Jr.[a], C.P. Cordero[a], H.M. Strebel[c], L. Palileo-Villanueva[b], J.V. Mantaring[a]

[a]*Department of Clinical Epidemiology, University of the Philippines College of Medicine, Manila, Philippines*
[b]*Division of Internal Medicine, University of the Philippines-Philippine General Hospital, Manila, Philippines*
[c]*Division of Medical Oncology, University of the Philippines-Philippine General Hospital, Manila, Philippines*

Accepted 12 November 2025; Published online 20 November 2025

Abstract

Background: Systematic reviews (SRs) with network meta-analyses (NMAs) are increasingly used to inform guidelines, health technology assessments (HTAs), and policy decisions. Their methodological complexity, as well as the difficulty in assessing the exchangeability assumption and the large amount of results, makes appraisal more challenging than for SRs with pairwise NMAs. Numerous SR- and NMA-specific appraisal tools exist, but they vary in scope, intended users, and methodological guidance, and few have been validated.

Objectives: To identify and describe appraisal instruments and interpretive guides for SRs and NMAs specifically, summarizing their characteristics, domain coverage, development methods, and measurement-property evaluations.

Methods: We conducted a methodological scoping review which included structured appraisal instruments or interpretive guides for SRs with or without NMA-specific domains, aimed at review authors, clinicians, guideline developers, or HTA assessors from published or gray literature in English. Searches (inception−August 2025) covered major databases, registries, organizational websites, and reference lists. Two reviewers independently screened records; data were extracted by one and checked by a second. We synthesized the findings narratively. First, we classified tools as either structured instruments or interpretive guides. Second, we grouped them according to their intended audience and scope. Third, we assessed available measurement-property data using relevant COnsensus-based Standards for the selection of health Measurement INstruments items.

Results: Thirty-four articles described 22 instruments (11 NMA-specific, nine systematic reviews with meta-analysis-specific, 2 encompassing both systematic reviews with meta-analysis and NMA). NMA tools added domains such as network geometry, transitivity, and coherence, but guidance on transitivity evaluation, publication bias, and ranking was either limited or ineffective. Reviewer-focused tools were structured with explicit response options, whereas clinician-oriented guides posed appraisal questions with explanations but no prescribed response. Nine instruments reported measurement-property data, with validity and reliability varying widely.

Conclusion: This first comprehensive map of systematic reviews with meta-analysis and NMA appraisal resources highlights the need for clearer operational criteria, structured decision rules, and integrated rater training to improve reliability and align foundational SR domains with NMA-specific content.

Keywords: Network meta-analysis; Systematic review; Critical appraisal tools; Methodological quality assessment; Reliability; Validity

1. Introduction

Systematic reviews (SRs) with network meta-analyses (NMAs) inform clinical practice guidelines (CPGs), health technology assessments (HTAs), and evidence-based decision-making [1,2], but pose appraisal challenges because of its unique assumptions and methodology [1,3].

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

* Corresponding author. Department of Clinical Epidemiology, University of the Philippines College of Medicine, Taft Avenue, Manila, Philippines 1000.

E-mail address: kmmondragon1@up.edu.ph (K.M. Mondragon).

https://doi.org/10.1016/j.jclinepi.2025.112056

Plain Language Summary

NMA is a way to compare many treatments at once by combining results from multiple studies—even when some treatments have not been directly compared head-to-head. Because NMAs are complex, users need clear tools to judge whether an analysis is trustworthy. We reviewed and mapped 22 instruments published over the last 3 decades that are used to appraise or interpret SRs and NMAs. About half were designed specifically for NMAs; the rest were general SR tools that are applicable to NMAs. Most tools cover the basics of good reviews (clear question, fair search, bias assessment, and transparent synthesis). NMA-specific tools also address issues unique to networks, such as how the network is connected, whether indirect and direct evidence agree (consistency), and how to interpret treatment rankings.

However, important gaps remain. Few tools give step-by-step checks for transitivity/consistency, network-level publication bias, or ranking uncertainty, and reported reliability between raters is inconsistent. Reporting checklists (eg, Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Network Meta-Analyses) specify what information should be reported but not how well it should be presented. Certainty frameworks (eg, Grading of Recommendations Assessment, Development, and Evaluation or Confidence in Network Meta-Analysis) outline how confidence in results is rated across domains such as inconsistency or imprecision, but they do not explain or standardize the different ways these domains are evaluated. What this means: guideline developers, HTA assessors, and clinicians should seek collaboration with statisticians experienced in NMA, and favor instruments with clear decision rules and user training. Better-tested, clearer tools will make NMA assessments more consistent and trustworthy.

Various instruments and user guides exist for SRs and NMAs [4,5]; however, evidence for their validation and usability is uneven. Although early reviews found that many appraisal tools lacked formal validation or reliability testing [6−8], subsequent instruments have undergone limited psychometric evaluation, and uncertainty persists about their comprehensiveness, reliability across users, and applicability in different review contexts [9,10].

No prior review has synthesized and compared the full spectrum of appraisal instruments and interpretive guidance across both NMA-specific and general SRs. Previous reviews have been narrower in scope [5,11,12], without cataloging the breadth of available tools for diverse users. The rapid expansion of NMAs in policy and guideline contexts [2] underscores the need for this updated and comprehensive synthesis.

General SR appraisal tools remain relevant because they are frequently applied to NMAs in practice and have provided the foundation for several NMA-specific instruments [13,14]. Mapping both together shows methodological continuity while clarifying which domains are unique to NMAs. This also highlights what SR tools omit, reinforcing the need for dedicated NMA appraisal guidance.

This scoping review addresses this gap by identifying, describing, and comparing appraisal frameworks over 3 decades, mapping content domains, and examining their development methods and intended end users, to inform the creation of validated and reliability-tested appraisal instruments.

2. Objectives

The primary objective of this methodological scoping review was to identify and characterize critical appraisal instruments and interpretive user guides applicable to SRs and NMAs specifically. Secondary objectives were to (1) describe each instrument's intended end-users, scope, structure, and content domains; (2) map coverage of foundational SR- and NMA-specific domains using a prespecified domain classification framework; and (3) summarize reported development methods and any empirical evaluation of measurement properties.

3. Materials and Methods

3.1. Study design

This study was conducted as a scoping review of critical-appraisal tools and user guides relevant to SRs and NMAs specifically, reported in accordance with Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) 2020.

3.2. Eligibility criteria

We included structured appraisal instruments and interpretive user guides with explicit criteria for appraising, interpreting, or reporting SRs with NMAs. General SR tools were eligible because NMAs are conducted within SR frameworks; aside from their common applicability to NMAs, some of these tools have provided the foundation for some NMA-specific instruments. Critical-appraisal tools were structured instruments with criteria/signaling questions assessing methodological quality, credibility, or risk of bias; user guides were interpretive resources aiding end-users' understanding or application of SR/NMA findings. Reporting checklists (eg, PRISMA-NMA) and certainty frameworks (eg, Grading of Recommendations Assessment, Development, and Evaluation for Network

What is new?

Key Findings

- This review mapped 22 appraisal resources across 3 decades.
- While general SR tools remain foundational and widely applied to NMAs, NMA-specific domains across different tools—particularly transitivity, coherence, network-level publication bias, and treatment ranking—vary in their definitions and in the level of guidance available.
- Only a few tools report reliability or validity testing, and observed agreement varies substantially, largely influenced by rater training and calibration.

What this adds to what is known?

- Previous methodological reviews were narrower in scope, but none comprehensively synthesized and compared appraisal or interpretive instruments for SRs and NMAs.
- This review integrates appraisal checklists, user guides, reporting guides and certainty frameworks to reveal fragmented coverage, limited psychometric evaluation, and absence of end-user calibration guidance across available instruments.

What is the implication and what should change now?

- Future tools should provide explicit operational criteria for transitivity, coherence, network-level publication bias, and treatment ranking.
- Embed reliability testing and rater training.
- Promote multidisciplinary appraisal teams including content experts and statisticians with NMA expertise to improve consistency, interpretability, and decision utility.

Meta-Analyses [GRADE-NMA], Confidence in Network Meta-Analysis [CINeMA]) because they constitute foundational frameworks that influence how NMAs are designed, reported, and interpreted. Their inclusion enables examination of how these established standards shape methodological expectations and interpretive practice among review authors and evidence assessors. Purely procedural conduct guidance without a checklist structure (eg, Cochrane Handbook and Cochrane Methodological Expectations of Cochrane Intervention Reviews) were excluded. All publication years and formats that addressed $\geq$1 prespecified appraisal domain were considered. Due to limited resources and the lack of validated, reproducible translations from automated tools (eg, DeepL) for methodological content, we included English-language sources only.

Tools were excluded if they lacked explicit SR/NMA guidance (eg, opinion pieces), addressed only other evidence types (eg, diagnostic-test or qualitative reviews), or were duplicate reports, in which case the most complete version was retained.

3.3. Search strategy and selection process

We searched MEDLINE (via PubMed), the Cochrane Methodology Register, Google Scholar, and CINAHL from inception to August 2025, supplemented by gray-literature sources (eg, PROSPERO, COnsensus-based Standards for the selection of health Measurement INstruments [COSMIN] database, OpenGrey, Critical Appraisal Skills Programme, Joanna Briggs Institute). Search strategies combined controlled vocabulary and free-text terms to appraisal, evidence synthesis, and NMA. Reference lists were screened for additional sources. The full search strategy for at least 1 database (eg, MEDLINE) and the complete database count and search terms are provided in the supplementary file (Tables S1 and S2).

Records were managed in Rayyan for identifying duplicates and facilitate screening. All removal of duplicates, inclusion and exclusion of studies were done manually. Titles, abstracts, and full texts were reviewed independently by two authors (K.M., R.V.) with disagreements resolved by consensus.

3.4. Data extraction

We piloted an Excel template and extracted information on tool identifiers, audience, purpose, development methods, structure, domains, links to frameworks, and reported measurement properties. Extraction was performed by 1 reviewer (K.M.) and checked by another (R.V.). Full extraction tables are provided in Supplementary S3-S6.

3.5. Data synthesis

We used a structured narrative synthesis, classifying resources as structured instruments or interpretive guides, and grouping them by audience (evidence reviewers vs clinicians) and scope (general SR vs NMA). Comparisons were made on purpose, development, structure, domain coverage, and measurement properties.

For structured instruments with empirical evaluation, we applied relevant COSMIN Risk of Bias items, focusing on content validity and interrater agreement [15]. COSMIN results were not aggregated into overall ratings. As no standard risk-of-bias tool exists for appraisal tool development, a formal RoB assessment was not performed. Missing information was coded as ‘‘not mentioned’’ or ‘‘not applicable.’’

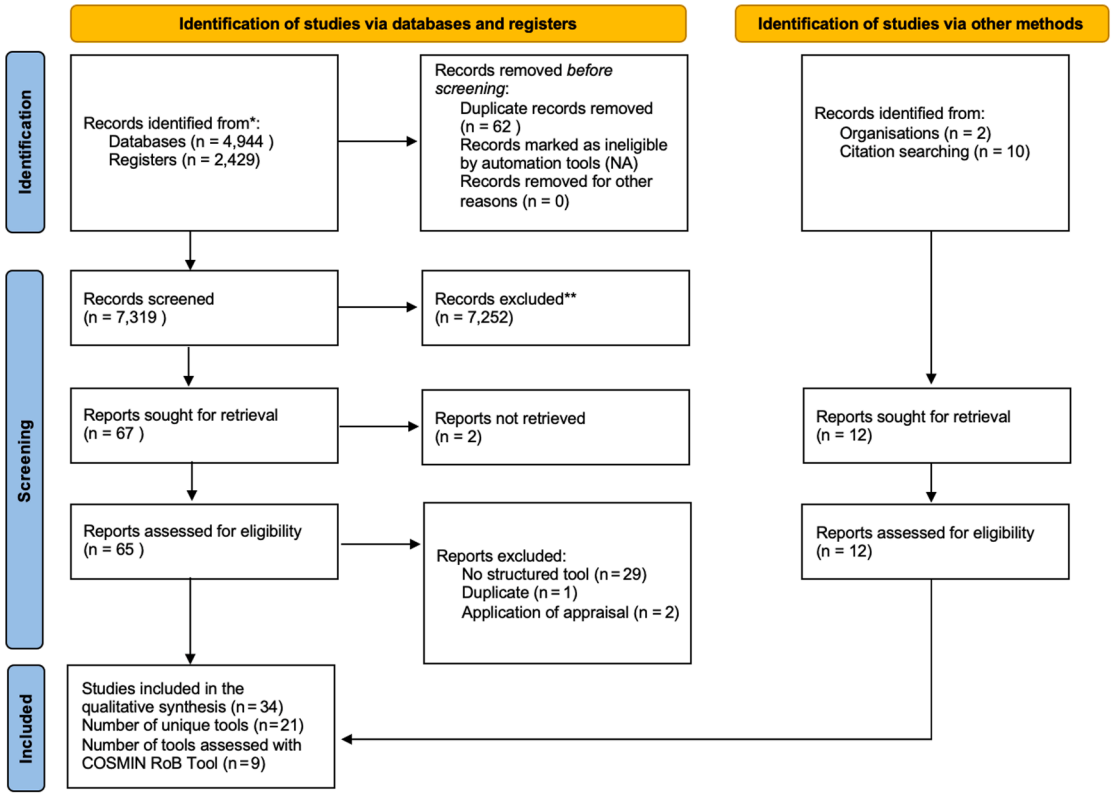

Figure. PRISMA flow diagram. COSMIN RoB Tool: COnsensus-based Standards for the selection of health Measurement INstruments Risk of Bias Tool; NA, not applicable; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

Findings were summarized in extraction tables and a domain mapping matrix.

4. Registration

This protocol is registered with PROSPERO (CRD420251089970) and Open Science Framework (DOI: 10.17605/OSF.IO/QX62 R).

5. Results

5.1. Overview of search and screening process

A comprehensive search strategy was implemented across PubMed, Cochrane Library, Google Scholar, and CINAHL with the last search conducted on August 11, 2025. We conducted a systematic search (see Supplementary Table S3); the number of records identified, screened, and included is shown in the PRISMA flow diagram [Fig]. We identified 34 studies reporting on 22 unique appraisal tools or frameworks. Included articles comprised original tool publications and their external validations.

5.2. Characteristics and structure of the tools

Of the 22 appraisal instruments identified, 11 were developed specifically for NMA [11,14,16–31], 9 for general SRs [13,32–42], and two comprehensive guides applicable to both pairwise meta-analysis and NMAs specifically, the *Evidence Synthesis for Decision Making: A Reviewers' Checklist* (2103) the *Users' Guides to the Medical Literature* (JAMA Users' Guides) [43,44]—which provide structured frameworks for evaluating and interpreting evidence syntheses (Table 1).

NMA-specific tools (2014–2019) ranged from structured checklists, such as the *Checklist for Critical Appraisal of Indirect Comparisons* (Ortega et al checklist) [16] and the *Critical Appraisal Tool for Network Meta-Analyses* (CAT-NMA, Lee unpublished doctoral thesis) [28], to narrative interpretive guides for clinicians, including the two-part *Network Meta-Analysis Users' Guide for Surgeons* by (Foote et al,; Chaudhry et al) [25,26], the *Users' Guides to the Medical Literature* by Guyatt et al (JAMA NMA guide) [44], the *Allergist's Guide to Network Meta-analysis* (Chu et al guide) [29], and the *Network Meta-Analysis User Guide for Pediatricians* by Al Khalifah et al (Al Khalifah et al pediatric guide) [27].

Table 1. Characteristics of included critical appraisal tools and user guides

Tool name	Year (latest)	Intended users	Target review type	NMA focus	Purpose	Validation/reliability evidence	Development method	Key notes/frameworks
Sacks' Checklist	1987	Unspecified	SRs, MAs	No	Appraisal tool	No	Consensus by the authors	Developed for use within the
OQAQ	1991	Readers, editors	SRs, MAs	No	Appraisal tool	Yes (face/construct validity, IRR)	Lit review, iteration	Early, influential
AMSTAR/AMSTAR-2	2007/2017	Evidence reviewers	SRs, MAs	No	Appraisal tool	Yes (content/ construct validity, IRR)	Lit review, expert consensus	Based on OQAQ, Sacks'
JBI's Checklist	2017	Evidence reviewers	SRs, MAs	No	Appraisal tool	No	Lit review, expert consensus	Widely used, not NMA-specific
CASP-SRMA	2024	Health professionals, evidence reviewers	SRs, MAs	No	Appraisal tool	No	Lit review, expert consensus	
ROBIS	2015	Review authors, guideline developers	SRs, MAs	No	Appraisal tool	Yes (external validation: IRR)	Lit review, consensus	Focus on risk of bias
SAQAT	2015	GRADE users	SRs, MAs	Partial	Certainty of evidence — semiauto (software)	Yes (IRR, Bayesian modeling)	Modeling, piloting	GRADE implementation
MECCIR	2024	Campbell authors, editors	SRs, MAs	No	Appraisal tool	No	Literature review; consensus	Conduct-reporting standards for Campbell reviews
CAT-HPPR	2022	Evidence reviewers	SRs, MAs	No	Appraisal tool	Yes (content and face validity) No IRR	Literature review, expert panel, piloting	
Evidence Synthesis Checklist (NICE)	2013	Evidence reviewers	SRs with MAs and NMA	Yes	Appraisal tool	Yes (content and face validity), no IRR	Expert led	Based on the technical support documents in evidence synthesis prepared for the National Institute for Health and Clinical Excellence (NICE) decision support unit.
Ortega et al. Checklist	2014	Drug reviewers	SRs with NMA	Yes	Appraisal tool	Yes (IRR)	Literature review, consensus	Simple, for indirect comparisons

(*Continued*)

Table 1. Continued

Tool name	Year (latest)	Intended users	Target review type	NMA focus	Purpose	Validation/reliability evidence	Development method	Key notes/frameworks
GRADE-NMA	2014/2023	Evidence reviewers	SRs with NMA	Yes	Certainty of evidence	No	Consensus, framework	Certainty rating for NMA
CINeMA	2014–2020	Evidence reviewers	SRs with NMA	Yes	Certainty of evidence	No	GRADE adaptation	Online GRADE for NMA
ISPOR NMA Task Force	2014	Evidence reviewers, HTA	SRs with NMA	Yes	Appraisal tool	Yes (% agreement)	Consensus, review	HTA-focused, structured guide
PRISMA-NMA	2015	Authors, peer reviewers	SRs with NMA	Yes	Reporting checklist	No (reporting focus)	Delphi, consensus	PRISMA extension for NMA
Checklist for Indirect Comparisons and NMAs	2015	End-users of NMAs	SRs with NMA	Yes	Narrative guide	No	Literature review	Review article
JAMA's Users' Guide for Medical Literature	2015	Clinicians educators	SRs with MAs and NMA	Yes	Interpretive guide	No	Expert-driven	Academic book with dedicated chapters for SRs and NMAs
NMA Users' Guide for Surgeons	2015	Surgeons	SRs with NMA	Yes	Interpretive guide	No	Expert-driven	JAMA's Users' Guide
NMA Users' Guide for Pediatricians	2018	Pediatricians	SRs with NMA	Yes	Interpretive guide	No	Expert-driven	JAMA's Users' Guide
CAT-NMA (Lee's Thesis)	2020 (unpublished)	Reviewers, educators	SRs with NMA	Yes	Appraisal tool	Yes (IRR)	Synthesis, expert input	Kappa IRR, academic thesis
NMA Users' Guide for allergists	2021	Allergists	SRs with NMA	Yes	Interpretive guide	No	Expert-driven	JAMA's Users' Guide
ROB-NMA Tool	2022–2025	Methodologists, guideline developers	SRs with NMA	Yes	Appraisal tool	Partial (content and face validity, piloting)	Lit review, consensus	17 bias domains

SRs, Systematic Reviews; MA, Meta-Analysis; NMA, Network Meta-Analysis; IRR: Interrater reliability; OQAQ, Overview of Quality Assessment Questionnaire; AMSTAR, A Measurement tool to assess SysTemAtic Reviews; ROBIS, Risk of BIas in Systematic Reviews; CAT-HPPR, Critical Appraisal Tool for Health Promotion and Prevention Reviews; CAT-NMA, Critical Appraisal Tool for Meta-analysis; MECCIR, Methodologic Expectations for Campbell Collaboration Intervention Reviews; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses; GRADE, Grading of Recommendations Assessment, Development, and Evaluation; CINeMA, Confidence In Network Meta-Analysis; ISPOR, International Society for Pharmacoeconomics and Outcomes Research; JAMA, Journal of the American Medical Association; HTA, health technology assessment; ROB-NMA, Risk-of-Bias in Network Meta-Analysis.

Validation/IRR: Indicates whether empirical validation (face/content/construct validity or interrater reliability) is reported in tool development.

NMA focus: "Yes" for tools developed or adapted specifically for network meta-analysis.

Early instruments such as the *Evidence Synthesis for Decision Making: A Reviewers' Checklist* (Ades et al) [43], *Indirect Treatment Comparison/Network Meta-Analysis Methodological Checklist* by Jansen et al (International Society for Pharmacoeconomics and Outcomes Research [ISPOR] tool) [22], Ortega et al checklist [16], and *Checklist for Evaluation of Indirect Comparisons and Network Meta-Analysis* (Kiefer et al checklist) [30] targeted evidence reviewers, combining SR foundational domains (eligibility criteria, search methods, bias assessment, synthesis, and results presentation) with NMA-specific domains (network geometry, transitivity, coherence, ranking). The PRISMA-NMA [23,24] introduced reporting standards but not appraisal criteria, yet its specification of key methodological and analytical items indirectly informed subsequent appraisal frameworks. Later tools, such as the GRADE-NMA [17,18] and CINeMA [19−21] operationalized certainty-of-evidence judgments across transitivity, incoherence, indirectness, imprecision, and publication bias. Although primarily designed for certainty appraisal, their structured domains underpin the credibility of NMA findings. The more recent Risk-of-Bias in Network Meta-Analysis (ROB-NMA) addresses NMA-specific risk of bias and complements SR appraisal tools [11,13,14,31].

General SR tools (1987−2017) included foundational checklists such as the *Sacks' Checklist* and the *Overview Quality Assessment Questionnaire* (OQAQ), which informed the development of *A MeaSurement Tool to Assess Systematic Reviews* (AMSTAR) and its update, *AMSTAR-2* [32−36,45,46]. The *Risk of Bias in Systematic Reviews* (ROBIS) tool shifted focus to risk of bias in SRs, using a domain-based approach and signaling questions [13]. The *Semi-Automated Quality Assessment Tool* (SAQAT), primarily based on the GRADE framework, incorporated Bayesian modeling to facilitate certainty judgments in a semiautomated tool online [39,40]. The *Critical Appraisal Tool for Health Promotion and Prevention Reviews* (CAT-HPPR) provided guidance for health policy reviewers [42]. The *Joanna Briggs Institute Critical Appraisal Checklist for Systematic Reviews and Research Syntheses* [37] and the *Critical Appraisal Skills Program Checklist for Systematic Reviews with Meta-Analysis* [38], offered interpretive frameworks with yes/no/unclear prompts but no formal psychometric validation. The *Methodologic Expectations for Campbell Collaboration Intervention Reviews* established methodological standards for Campbell SRs [41].

Foundational evidence synthesis domains, such as research question framing, eligibility criteria, search methods, study selection, bias assessment, synthesis approaches, and results presentation, were covered by most systematic reviews with meta-analysis-oriented tools (eg, OQAQ, AMSTAR-2, ROBIS, CAT-HPPR) [13,33−36,42,43,46] and some clinician-oriented NMA tools [25−27,29,44].

NMA-specific instruments incorporated five additional domains: network geometry, transitivity, consistency, indirectness, and treatment ranking [14,16−27,29,30,43,44], with all covering transitivity and coherence [14,16−27,29,30,43,44], and most (72%) evaluated network geometry [14,16−29,43,44]. However, depth and clarity of operationalization varied markedly (Supplementary Table S4). Only 27% of NMA tools provide structured procedures for detecting network-level publication bias [14,17−21], fewer than one-third offer stepwise criteria for transitivity, incoherence, or indirectness appraisal [14,16−22,43], and only a minority formally guide treatment-ranking interpretation [22,26,27,29,44] (Supplementary Table S5).

Evidence reviewer−focused tools used explicit criteria and decision rules [14,16−22,43], whereas clinician-oriented and teaching-oriented guides relied on illustrative examples [25−27,29,44].

5.3. Development methods and validation

Clinician-oriented guides were primarily narrative teaching resources, developed by small author teams with subject-matter expertise. These guides support the interpretation of NMA findings in specific clinical contexts (eg, surgery, pediatrics, allergy) and were presented as text-based explanations, illustrative examples, and interpretive checklists without explicit scoring. Development was informal, relying on the authors' synthesis of literature and experience, and did not include systematic field testing or psychometric evaluation to assess reliability or validity [25−27,29,44]. In contrast, reviewer-focused tools are structured appraisal instruments for SR authors, and guideline/HTA assessors. Their development followed multistage, transparent methods, beginning with a review of existing instruments to identify relevant domains, followed by consensus-building techniques, and culminating in pilot testing for usability and clarity [13,14,22,34,36,42,43]. For example, AMSTAR 2 underwent multiple Delphi rounds involving experts and then interrater reliability testing [34]. The ISPOR questionnaire employed stakeholder consultation and was refined through user-testing before publication [22].

5.4. COSMIN RoB assessments

Nine instruments had empirical measurement-property data. For the COSMIN RoB assessment, AMSTAR and AMSTAR 2 were appraised separately due to their substantive differences, yielding nine instruments in total, though they were counted as a single instrument in the overall synthesis. Among the general SR tools, OQAQ demonstrated strong reliability (intraclass correlation coefficient [ICC] = 0.91 across 36 reviews with nine raters) and construct validity, while AMSTAR showed acceptable agreement and AMSTAR 2 expanded coverage to non-randomized studies, demonstrating moderate reliability (median $\kappa \approx 0.51$, range 0.15−0.80; percent agreement up

Table 2. Measurement properties of appraisal tools for systematic reviews and network meta-analyses

Tool	Study sample	Validity evidence	Interrater reliability
AMSTAR [25]	2 raters; 42 SRs [25] 4 raters, 30 SRs [44]	Convergent validity with global assessment (r = 0.72); content validity via expert consensus and factor analysis; no criterion validity [25]	Item κ = 0.34–0.80; 9 items κ ≥ 0.75; ICC = 0.84 (total score); Pearson's r = 0.96 [25]; κ range = 0.16-0.88 with a median of 0.62 (substantial agreement) [31]
AMSTAR-2 [24]	6 raters, 54 SRs [24]; external validation: 4 raters, 60 SRs [44]	Content validity via expert consensus; convergent validity with AMSTAR (r = 0.91) and ROBIS (r = 0.84) [44]	Median κ = 0.51 (item range 0.15 –0.80) [44]; Moderate or better agreement per-item 92% [24]
ROBIS [9]	Delphi: ~29 raters per round; 3 pairs of independent reviewers: 8 articles [9]; external validation: 4 raters, 30 SRs [44]	Content validity via Delphi consensus; relevance and comprehensibility implied; no criterion validity [9]	Median κ = 0.27 (range −0.01 to 0.54) (based on 4 raters) [44]; domain-level consensus 65% (based on 1 pair of reviewers) [9]
ISPOR [15]	~18 raters; 3 NMA articles [15]	Content validity via concept elicitation and consensus; comprehensibility assessed in pilot [15]	% agreement average 72% (range 42% –91%); no κ/ICC reported [15]
OQAQ [23,42]	9 raters; 36 reviews [23,42]	Construct validity (6/7 a priori hypotheses confirmed); criterion validity (correlated with self-reported rigor); face validity tested with clinicians [23,42]	ICC >0.5 for most items; 60% > 0.7; weaker agreement for subjective items [23,42]
Ortega Checklist [16]	72 raters (pharmacists, drug evaluation context) [16]	Content validity from literature/expert consensus; external review confirmed relevance/comprehensibility [16]	κ = 0.83 (quality domain), 0.61 (clinical), 0.63 (methods/stats) [16]
Lee's CAT-NMA [21]	7 raters; 3 NMAs [21]	Face/content validity from adaptation of existing tools and expert input [21]	Fleiss' κ range 0.029–0.643 (slight –substantial, most items slight–fair) [21]
RoB NMA Tool [7,10,39,40]	Delphi (28 in round 1, 22 in round 2); knowledge user survey (298); pilot (21 raters) [7,10,39,40]	Strong content and face validity: systematic item generation, Delphi, stakeholder survey, pilot testing [7,10,39,40]	No published κ/ICC yet; pilot tested feasibility, but time burden high (median 79 min) [7,10,39,40]
SAQAT [28,29]	2 raters; 28 meta-analyses; 15 Cochrane reviews [28,29]	Content validity via alignment with GRADE and iterative refinement; partial criterion validity vs manual GRADE [28,29]	κ = 0.63–1.00 (per domain), κ = 0.79 (overall quality); with GRADE, initial κ = 0.11 → improved κ = 0.35 after amendments [28,29]

AMSTAR, A Measurement Tool to Assess systematic Reviews; ROBIS, Risk Of Bias in Systematic reviews; ISPOR, International Society for Pharmacoeconomics and Outcomes Research; ICC, intraclass correlation coefficient; OQAQ, Overview Quality Assessment Questionnaire; RoB, Risk of Bias; NMA, Network Meta-Analysis; SAQAT, Semi-Automated Quality Assessment Tool; SRs, Systematic Reviews; GRADE, Grading of Recommendations Assessment, Development, and Evaluation system.

to 92% across 60 reviews) and strong convergent validity [10,33,34,46]. ROBIS showed lower interrater reliability in an external validation (median κ ≈ 0.27) [10,13]. Certainty-rating tools such as SAQAT achieved substantial agreement (κ ≈ 0.79 overall, κ = 0.63–1.00 across domains) and validated its Bayesian model approach against manual GRADE [39,40].

NMA-specific tools showed variable results. Ortega et al's checklist achieved high kappas in its domains (κ = 0.61–0.83 across 72 pharmacist raters), while Lee's CAT-NMA showed only slight to fair agreement for most items (Fleiss' κ = 0.03–0.64 across seven raters and three NMAs) [16,28]. The ISPOR questionnaire reported raw agreement averaging 72% (range 42%–91%) among 22 raters [22]. The RoB-NMA tool underwent extensive development, including Delphi rounds and a large stakeholder survey, but has yet to publish formal reliability testing [10,41,42]. Overall, across both systematic reviews with meta-analysis and NMA instruments, face and content validity were consistently supported, but psychometric testing beyond interrater agreement was rare. Instruments also varied widely in rater numbers, sample sizes, and whether calibration phases were undertaken (Table 2).

6. Discussion

Our review mapped the current landscape of appraisal resources, encompassing 22 distinct instruments. These ranged from highly specialized tools, 11 developed specifically for NMAs, 9 for general SRs, and two broad, well-established references, the JAMA Users' Guides and the Evidence Synthesis for Decision Making. This breadth reflects the diversity of approaches available to support end-users in assessing evidence across different types of syntheses. The scope and depth of these instruments varied substantially, reflecting differences in target audience, methodological rigor during development, and degree of empirical testing.

This review synthesized appraisal instruments and interpretive user guides for SRs with or without NMA, spanning publications from 1987 to 2019. The timeline highlights a notable expansion of NMA-specific resources after 2015, coinciding with both the formalization of reporting guidance [23] and the increasing uptake of NMA in the medical literature [2]. Although the statistical assumptions underpinning NMAs have remained constant [14,16−21,23,25−27,29,30,44], methodologies have expanded to include more treatment-ranking metrics [47,48], network-level publication bias [46], and Bayesian modeling approaches [13,14,23].

However, appraisal tools have not consistently kept pace with these methodological advances. Updated frameworks such as GRADE-NMA [17,18,49,50], CINeMA [19−21], and the PRISMA-NMA extension [23,24] primarily target review authors, focusing on certainty of evidence, reporting, and analytic transparency. On the end-user side, although recent instruments such as the ROB-NMA tool [14] represent important progress, the overall range of resources for guideline developers, HTA assessors, and clinicians remains limited, with many narrative guides retaining their original scope despite substantial advances in NMA methodology and standards [25−27,29].

6.1. Operational gaps in NMA-specific domains

All structured NMA tools addressed transitivity and coherence, yet the operationalization of these domains varied. Few tools provided prespecified diagnostic steps or thresholds for judging transitivity/incoherence; network-level publication bias assessment was rarely operationalized; and treatment ranking credibility was often mentioned without detailed guidance. These gaps echo observations in the broader methodological literature, which has highlighted the challenges of operationalizing assumptions such as transitivity and consistency [51−54] and the limitations on interpretation of network-level publication bias and ranking [55,56]. This lack of operational clarity may increase subjectivity and contribute to lower reliability, particularly when tools are applied without training.

6.2. Reliability patterns and contributing factors

Among the nine instruments with empirical measurement-property data, interrater reliability varied widely (ICC 0.91 for OQAQ [33] to κ as low as 0.03 for some CAT-NMA items [28]). This variability was not solely attributable to the instrument's structure but also to the methodology of its application.

Older instruments such as the OQAQ, which relied on simple yes/no items, yielded higher agreement (ICC ≈ 0.91) [33,46] by limiting subjectivity and variability in interpretation. In contrast, more recent tools (eg, ROBIS [13] and CAT-NMA [28]) adopt nuanced response scales (eg, "probably yes," "probably no") and target more complex constructs. While this improves sensitivity and contextual relevance, it broadens rater interpretation, and lowers observed agreement.

Rater preparation was another key determinant of reliability. Higher κ or ICC values were generally observed in studies with structured training and calibration before formal reliability assessment (eg, AMSTAR-2 [24,25], OQAQ [23,44]). Whereas minimal or no training (eg, ROBIS [10,13], CAT-NMA [28]) demonstrated lower reliability. For example, ROBIS, dropped κ to ≈0.27 in external validation when raters had limited prior exposure compared with AMSTAR/AMSTAR-2 [10,13].

Expansion of appraisal scope in NMA-specific tools (eg, network geometry, transitivity, and advanced modeling) also demands greater statistical expertise. Without explicit operational guidance, consistent scoring becomes challenging. Improving reliability will require refining item wording, structured decision rules, and systematic rater calibration, particularly for complex, multidomain assessments.

6.3. Methodological gaps and future priorities

Several priorities for future tool development emerge from these findings. First, operational detail is needed for key NMA domains, particularly for transitivity and incoherence, recognizing network-level bias, and treatment-ranking uncertainty assessment. Second, reliability can be improved through user-tested decision rules: piloting items with target users, identifying areas of frequent disagreement, and refining wording/structure accordingly. Third, reporting of reliability metrics should extend beyond raw agreement to include κ or ICC estimates, ideally based on independent appraisal of real NMAs under realistic use conditions. Fourth, bridging foundational SR domains with NMA-specific constructs would enhance methodological coherence and completeness across tools. Finally, given the technical sophistication of NMA statistical models, active collaboration with statisticians experienced in evidence synthesis is essential. Tools should therefore encourage or explicitly require multidisciplinary appraisal teams that include both content and methodological experts to minimize misapplication and enhance credibility.

6.4. Methodological considerations and limitations

This review has several limitations. We compared tools with different purposes (appraisal checklists, reporting standards, certainty frameworks), which limited direct comparability. Measurement-property studies were heterogeneous in design and metrics (κ, ICC, raw agreement). Because COSMIN was developed for health measurement instruments, we applied it only narratively for validity and reliability; other domains were not applicable. Assigning items to ''foundational'' or ''NMA-specific'' domains required judgment, as some concepts spanned both. Finally, we did not assess usability, time burden, or whether tool use changes downstream decisions.

7. Conclusion

This review provides the first comprehensive mapping of appraisal resources for systematic reviews with meta-analysis and NMAs, identifying major gaps and priorities for tool development. Despite rapid expansion, several domains remain underoperationalized and reliability is inconsistent. Advancing future instruments will require clearer decision rules, structured rater training, and stronger alignment between SR foundational domains and NMA-specific content. Equally essential is collaboration with statisticians experienced in NMA methods to ensure appropriate application and interpretation of complex models. Without such expertise, the expanding NMA toolkit risks misapplication and misinterpretation, compromising validity. Strengthening methodological rigor and interdisciplinary collaboration will enhance the consistency, transparency, and decision utility of NMA appraisal in research, and policy contexts.

Ethics statement

Ethical approval was not required for this methodological review.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT to assist with language editing. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

CRediT authorship contribution statement

K.M. Mondragon: Writing − review & editing, Writing − original draft, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **C.S. Tan-Lim:** Writing − review & editing, Supervision, Methodology, Investigation, Formal analysis, Data curation. **R. Velasco:** Writing − review & editing, Investigation, Formal analysis, Data curation. **C.P. Cordero:** Writing − review & editing, Supervision, Methodology, Formal analysis, Data curation. **H.M. Strebel:** Writing − review & editing, Supervision, Methodology, Data curation. **L. Palileo-Villanueva:** Writing − review & editing, Supervision, Methodology, Formal analysis, Data curation, Conceptualization. **J.V. Mantaring:** Writing − review & editing, Supervision, Methodology, Data curation, Conceptualization.

Declaration of competing interest

There are no competing interests for any author.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2025.112056.

Data availability

All data extracted and analyzed in this review are included within the article and its supplementary materials.

References

[1] Watt J, Tricco AC, Straus S, Veroniki AA, Naglie G, Drucker AM. Research techniques made simple: network meta-analysis. J Invest Dermatol 2019;139:4−12.e1. https://doi.org/10.1016/j.jid.2018.10.028.
[2] Zarin W, Veroniki AA, Nincic V, Vafaei A, Reynen E, Motiwala SS, et al. Characteristics and knowledge synthesis approach for 456 network meta-analyses: a scoping review. BMC Med 2017;15:3. https://doi.org/10.1186/s12916-016-0764-6.
[3] Ades AE, Welton NJ, Dias S, Phillippo DM, Caldwell DM. Twenty years of network meta-analysis: continuing controversies and recent developments. Res Synth Methods 2024;15:702−27. https://doi.org/10.1002/jrsm.1700.
[4] Shaheen N, Shaheen A, Ramadan A, Hefnawy MT, Ramadan A, Ibrahim IA, et al. Appraising systematic reviews: a comprehensive guide to ensuring validity and reliability. Front Res Metr Anal 2023;8:1268045. https://doi.org/10.3389/frma.2023.1268045.
[5] Swanepoel L-M, Brand A, Lourens A, Schoonees A, McCaul M. Methods resources for authors new to conducting systematic reviews with network meta-analysis: a scoping review. J Clin Epidemiol 2025;182:111759. https://doi.org/10.1016/j.jclinepi.2025.111759.
[6] Katrak P, Bialocerkowski AE, Massy-Westropp N, Kumar VS, Grimmer KA. A systematic review of the content of critical appraisal tools. BMC Med Res Methodol 2004;4:22. https://doi.org/10.1186/1471-2288-4-22.
[7] Burda BU, Holmer HK, Norris SL. Limitations of A measurement tool to assess systematic reviews (AMSTAR) and suggestions for improvement. Syst Rev 2016;5:58. https://doi.org/10.1186/s13643-016-0237-1.
[8] Pieper D, Mathes T, Eikermann M. Impact of choice of quality appraisal tool for systematic reviews in overviews. J Evid Based Med 2014;7:72−8. https://doi.org/10.1111/jebm.12097.

[9] Perry R, Whitmarsh A, Leach V, Davies P. A comparison of two assessment tools used in overviews of systematic reviews: ROBIS versus AMSTAR-2. Syst Rev 2021;10:273. https://doi.org/10.1186/s13643-021-01819-x.

[10] Lorenz RC, Matthias K, Pieper D, Wegewitz U, Morche J, Nocon M, et al. A psychometric study found AMSTAR 2 to be a valid and moderately reliable appraisal tool. J Clin Epidemiol 2019;114:133−40. https://doi.org/10.1016/j.jclinepi.2019.05.028.

[11] Lunny C, Veroniki A, Higgins JPT, Dias S, Hutton B, Wright JM, et al. Methodological review of NMA bias concepts provides groundwork for the development of a list of concepts for potential inclusion in a new risk of bias tool for network meta-analysis (RoB NMA Tool). Syst Rev 2024;13:25. https://doi.org/10.1186/s13643-023-02388-x.

[12] Guo JD, Gehchan A, Hartzema A. Selection of indirect treatment comparisons for health technology assessments: a practical guide for health economics and outcomes research scientists and clinicians. BMJ Open 2025;15:e091961. https://doi.org/10.1136/bmjopen-2024-091961.

[13] Whiting P, Savović J, Higgins JPT, Caldwell DM, Reeves BC, Shea B, et al. ROBIS: a new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol 2016;69:225−34. https://doi.org/10.1016/j.jclinepi.2015.06.005.

[14] Lunny C, Higgins JPT, White IR, Dias S, Hutton B, Wright JM, et al. Risk of bias in network meta-analysis (RoB NMA) tool. BMJ 2025;388:e079839. https://doi.org/10.1136/bmj-2024-079839.

[15] Mokkink LB, Elsman EBM, Terwee CB. COSMIN guideline for systematic reviews of patient-reported outcome measures version 2.0. Qual Life Res 2024;33:2929−39. https://doi.org/10.1007/s11136-024-03761-6.

[16] Ortega A, Fraga MD, Alegre-del-Rey EJ, Puigventós-Latorre F, Porta A, Ventayol P, et al. A checklist for critical appraisal of indirect comparisons. Int J Clin Pract 2014;68:1181−9. https://doi.org/10.1111/ijcp.12487.

[17] Puhan MA, Schünemann HJ, Murad MH, Li T, Brignardello-Petersen R, Singh JA, et al. A GRADE working group approach for rating the quality of treatment effect estimates from network meta-analysis. BMJ 2014;349:g5630. https://doi.org/10.1136/bmj.g5630.

[18] Izcovich A, Chu DK, Mustafa RA, Guyatt G, Brignardello-Petersen R. A guide and pragmatic considerations for applying GRADE to network meta-analysis. BMJ 2023;381:e074495. https://doi.org/10.1136/bmj-2022-074495.

[19] Salanti G, Giovane CD, Chaimani A, Caldwell DM, Higgins JPT. Evaluating the quality of evidence from a network meta-analysis. PLoS One 2014;9:e99682. https://doi.org/10.1371/journal.pone.0099682.

[20] Papakonstantinou T, Nikolakopoulou A, Higgins JPT, Egger M, Salanti G. CINeMA: software for semiautomated assessment of the confidence in the results of network meta-analysis. Campbell Syst Rev 2020;16:e1080. https://doi.org/10.1002/cl2.1080.

[21] Nikolakopoulou A, Higgins JPT, Papakonstantinou T, Chaimani A, Giovane CD, Egger M, et al. CINeMA: an approach for assessing confidence in the results of a network meta-analysis. PLoS Med, San Francisco, California, USA. 2020;17:e1003082. https://doi.org/10.1371/journal.pmed.1003082.

[22] Jansen JP, Trikalinos T, Cappelleri JC, Daw J, Andes S, Eldessouki R, et al. Indirect treatment comparison/network meta-analysis study questionnaire to assess relevance and credibility to inform health care decision making: an ISPOR-AMCP-NPC good practice task force report. Value Health 2014;17:157−73. https://doi.org/10.1016/j.jval.2014.01.004.

[23] Hutton B, Salanti G, Caldwell DM, Chaimani A, Schmid CH, Cameron C, et al. The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. Ann Intern Med 2015;162:777−84. https://doi.org/10.7326/M14-2385.

[24] Veroniki AA, Hutton B, Stevens A, McKenzie JE, Page MJ, Moher D, et al. Update to the PRISMA guidelines for network meta-analyses and scoping reviews and development of guidelines for rapid reviews: a scoping review protocol. JBI Evid Synth 2025;23:517−26. https://doi.org/10.11124/JBIES-24-00308.

[25] Foote CJ, Chaudhry H, Bhandari M, Thabane L, Furukawa TA, Petrisor B, et al. Network meta-analysis: users' guide for surgeons: part I − credibility. Clin Orthop Relat Res 2015;473:2166−71. https://doi.org/10.1007/s11999-015-4286-x.

[26] Chaudhry H, Foote CJ, Guyatt G, Thabane L, Furukawa TA, Petrisor B, et al. Network meta-analysis: users' guide for surgeons: part II - certainty. Clin Orthop Relat Res 2015;473:2172. https://doi.org/10.1007/s11999-015-4287-9.

[27] Al Khalifah R, Florez ID, Guyatt G, Thabane L. Network meta-analysis: users' guide for pediatricians. BMC Pediatr 2018;18:180. https://doi.org/10.1186/s12887-018-1132-9.

[28] Lee A. Developing critical appraisal of systematic reviews reporting network meta-analysis. University of Oxford; 2020.

[29] Chu D, Brignardello-Petersen R, Guyatt GH, Ricci C, Genuneit J. Method's corner: allergist's guide to network meta-analysis. Pediatr Allergy Immunol 2022;33:e13609. https://doi.org/10.1111/pai.13609.

[30] Kiefer C, Sturtz S, Bender R. Indirect comparisons and network meta-analyses. Dtsch Arztebl Int 2015;112:803−8. https://doi.org/10.3238/arztebl.2015.0803.

[31] Lunny C, Veroniki AA, Hutton B, White I, Higgins J, Wright JM, et al. Knowledge user survey and Delphi process to inform development of a new risk of bias tool to assess systematic reviews with network meta-analysis (RoB NMA tool). BMJ EBM 2023;28:58−67. https://doi.org/10.1136/bmjebm-2022-111944.

[32] Sacks H, Berrier J, Reitman D, Ancona-Berk, Chalmers T. Meta-analysis of randomized controlled trials. New Engl J Med 1987;316:450−5.

[33] Oxman AD, Guyatt GH. Validation of an index of the quality of review articles. J Clin Epidemiol 1991;44:1271−8. https://doi.org/10.1016/0895-4356(91)90160-B.

[34] 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. https://doi.org/10.1136/bmj.j4008.

[35] Shea BJ, Bouter LM, Peterson J, Boers M, Andersson N, Ortiz Z, et al. External validation of a measurement tool to assess systematic reviews (AMSTAR). PLoS One 2007;2:e1350. https://doi.org/10.1371/journal.pone.0001350.

[36] Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol 2007;7:10. https://doi.org/10.1186/1471-2288-7-10.

[37] Aromataris E, Fernandez R, Godfrey CM, Holly C, Khalil H, Tungpunkom P. Summarizing systematic reviews: methodological development, conduct and reporting of an umbrella review approach. Int J Evid Based Healthc 2015;13:132−40. https://doi.org/10.1097/XEB.0000000000000055.

[38] Critical Appraisal Skills Programme. CASP Systematic Review with Meta-Analysis Checklist. Oxford, Oxfordshire, United Kingdom: Critical Appraisal Skilles Programme; 2024. Available at: https://casp-uk.net/casp-tools-checklists. Accessed December 5, 2025.

[39] Stewart GB, Higgins JPT, Schünemann H, Meader N. The use of Bayesian networks to assess the quality of evidence from research synthesis: 1. PLoS One 2015;10:e0114497. https://doi.org/10.1371/journal.pone.0114497.

[40] Llewellyn A, Whittington C, Stewart G, Higgins JP, Meader N. The use of Bayesian networks to assess the quality of evidence from research synthesis: 2. inter-rater reliability and comparison with standard GRADE assessment. PLoS One 2015;10:e0123511. https://doi.org/10.1371/journal.pone.0123511.

[41] Aloe AM, Dewidar O, Hennessy EA, Pigott T, Stewart G, Welch V, et al. Campbell standards: modernizing Campbell's methodologic expectations for Campbell collaboration intervention reviews (MECCIR). Campbell Syst Rev 2024;20:e1445. https://doi.org/10.1002/cl2.1445.

[42] Heise TL, Seidler A, Girbig M, Freiberg A, Alayli A, Fischer M, et al. CAT HPPR: a critical appraisal tool to assess the quality of systematic, rapid, and scoping reviews investigating interventions in health promotion and prevention. BMC Med Res Methodol 2022; 22:334. https://doi.org/10.1186/s12874-022-01821-4.

[43] Ades AE, Caldwell DM, Reken S, Welton NJ, Sutton AJ, Dias S. Evidence synthesis for decision making 7. Med Decis Making 2013;33: 679−91. https://doi.org/10.1177/0272989X13485156.

[44] Guyatt G, Rennie D, Meade M, Cook D. Users' guides to the medical literature: essentials of evidence-based clinical practice. 3rd ed. Chicago, IL: Journal of the American Medical Association; 2015.

[45] Shea B, Boers M, Grimshaw JM, Hamel C, Bouter LM. Does updating improve the methodological and reporting quality of systematic reviews? BMC Med Res Methodol 2006;6:27. https://doi.org/10.1186/1471-2288-6-27.

[46] Oxman AD, Guyatt GH, Singer J, Goldsmith CH, Hutchison BG, Milner RA, et al. Agreement among reviewers of review articles. J Clin Epidemiol 1991;44:91−8. https://doi.org/10.1016/0895-4356(91)90205-N.

[47] Salanti G, Nikolakopoulou A, Efthimiou O, Mavridis D, Egger M, White IR. Introducing the treatment hierarchy question in network meta-analysis. Am J Epidemiol 2021;191:930−8. https://doi.org/10.1093/aje/kwab278.

[48] Rücker G, Schwarzer G. Ranking treatments in frequentist network meta-analysis works without resampling methods. BMC Med Res Methodol 2015;15:58. https://doi.org/10.1186/s12874-015-0060-8.

[49] Brignardello-Petersen R, Bonner A, Alexander PE, Siemieniuk RA, Furukawa TA, Rochwerg B, et al. Advances in the GRADE approach to rate the certainty in estimates from a network meta-analysis. J Clin Epidemiol 2018;93:36−44. https://doi.org/10.1016/j.jclinepi.2017.10.005.

[50] Brignardello-Petersen R, Izcovich A, Rochwerg B, Florez ID, Hazlewood G, Alhazanni W, et al. GRADE approach to drawing conclusions from a network meta-analysis using a partially contextualised framework. BMJ 2020;371:m3907. https://doi.org/10.1136/bmj.m3907.

[51] Veroniki AA, Vasiliadis HS, Higgins JPT, Salanti G. Evaluation of inconsistency in networks of interventions. Int J Epidemiol 2013; 42:332−45. https://doi.org/10.1093/ije/dys222.

[52] Dias S, Welton NJ, Caldwell DM, Ades AE. Checking consistency in mixed treatment comparison meta-analysis. Stat Med 2010;29: 932−44. https://doi.org/10.1002/sim.3767.

[53] Spineli LM. An empirical study on 209 networks of treatments revealed intransitivity to be common and multiple statistical tests suboptimal to assess transitivity. BMC Med Res Methodol 2024;24:301. https://doi.org/10.1186/s12874-024-02436-7.

[54] Spineli LM, Papadimitropoulou K, Kalyvas C. Exploring the transitivity assumption in network meta-analysis: a novel approach and its implications. Stat Med 2025;44:e70068. https://doi.org/10.1002/sim.70068.

[55] Chaimani A, Higgins JPT, Mavridis D, Spyridonos P, Salanti G. Graphical tools for network meta-analysis in STATA. PLoS One 2013;8:e76654. https://doi.org/10.1371/journal.pone.0076654.

[56] Salanti G, Ades AE, Ioannidis JPA. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. J Clin Epidemiol 2011;64:163−71. https://doi.org/10.1016/j.jclinepi.2010.03.016.