Evaluating the Quality of Systematic Reviews: A Narrative Review of Current Appraisal Frameworks and Introduction of the High Yield Med Tool

Nouran Alwisi[1], Fatima R. Alsharif[1], and Ayman Musleh*[2]

[1]College of Medicine, QU Health, Qatar University, Doha, Qatar 2. Department of Ophthalmology, Eye Specialty Hospital, Amman, Jordan

Keywords: Artificial Intelligence, Systematic reviews, High Yield Med, Quality Assessment

DOI: 10.59707/hymrOPKJ4706

Published on: December 1, 2025

Abstract

AI-assisted workflows are transforming the way systematic reviews are conducted, converting complex evidence synthesis processes into rapid, high-throughput outputs. This shift significantly reduces the time required for evidence synthesis, making systematic reviews more scalable and accessible. A critical comparison of existing appraisal tools including AMSTAR/AMSTAR 2, ROBIS, JBI, CASP, MECIR, and GRADE highlights that these frameworks focus primarily on transparent reporting and retrospective methodological quality, failing to capture the integrity of the review process, measure reproducibility, or adequately assess automation checkpoints inherent in modern hybrid workflows. To address this gap and support critical appraisal, we introduce the High Yield Med Quality Evaluation Tool (HYMQET), a novel framework designed to provide a structured, quantitative assessment of workflow quality, methodological rigor, and automation transparency in both human-led and hybrid human-AI systematic reviews. The HYMQET employs a stepwise, workflow-based scoring system across five core domains: Query Development, Screening Quality, Field Selection for Data Extraction, Full-Text Data Extraction, and Manuscript Writing. Its quantitative, workflow-based structure makes it an essential tool for the external validation, quality control, and reliable benchmarking of emerging automated and hybrid systematic review methodologies.

Introduction

Systematic reviews are the cornerstone of evidence-based medicine, providing a comprehensive synthesis of data from primary studies into coherent and clinically meaningful conclusions [1]. Their impact, however, is contingent on methodological rigor and transparency [2]. Adherence to standardised procedures, ranging from literature review to risk-of-bias assessment, is essential to ensure validity and reproducibility and any deviation from these established methods can lead to selective reporting and compromised reliability. The conclusions of systematic reviews directly inform clinical guidelines and health policy; therefore maintaining a verifiable and transparent workflow is paramount. The emer-

*Corresponding author: Ayman Musleh ¡aimanmesleh@gmail.com¿

gence of AI-assisted workflows, including automated screening algorithms and large language models (LLMs), has transformed how evidence is synthesised. These systems improve efficiency and reduce manual workload but are not without their limitations. Assessing review quality now requires evaluation of both the algorithmic contribution, and the adequacy of human oversight throughout the process. Existing appraisal tools were not designed to capture these evolving methodological dimensions[3,4]. Existing tools, including Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [5], A MeaSurement Tool to Assess Systematic Reviews (AMSTAR2) [6], and Risk of Bias In Systematic reviews (ROBIS) [7], have strengthened the standardisation of reporting and methodological assessment in systematic reviews. PRISMA focuses on transparent reporting, while AMSTAR2 and ROBIS evaluate methodological rigor at the review level. These tools, however, primarily address the output of the review rather than the quality of its underlying process. They do not quantify the consistency with which methods are applied, the handling of deviations from standard protocol or document the overall workflow. With the growing integration of AI-assisted workflows, these limitations become increasingly evident. To address this gap, the present review examines the evolution of systematic review appraisal tools and introduces the High Yield Med Evaluation Framework. This model provides a structured approach for assessing workflow quality, methodological rigor and transparency across both human-led and AI-assisted systematic reviews.

Overview of Existing Systematic Review Appraisal Tools

AMSTAR and AMSTAR 2

AMSTAR [8], first published in 2007 and later updated as AMSTAR 2 in 2017 [6], remains one of the most widely used tools for assessing the methodological quality of systematic reviews of interventions. The original AMSTAR provided an 11-item checklist assessing key domains such as protocol registration, search strategy, data extraction, and synthesis methods. AMSTAR 2 expanded this framework to 16 items, addressing additional aspects including non-randomised studies and the assessment of risk of bias in individual studies. Unlike its predecessor, AMSTAR 2 discourages the use of an overall numerical score, favouring a domain-based rating of confidence in results ("high," "moderate," "low," or "critically low"). This approach improves conceptual rigor but limits direct comparability between reviews. Nonetheless, AMSTAR 2 remains the gold standard for assessing methodological robustness in reviews of interventions.

ROBIS (Risk of Bias in Systematic Reviews)

The ROBIS tool [7], introduced in 2016, was developed to evaluate bias specifically at the review level, rather than the primary study level. It is structured into three phases: assessing relevance, identifying concerns across four domains: study eligibility criteria, identification and selection of studies, data collection and appraisal, and synthesis and findings, and determining the overall risk of bias. ROBIS is particularly suited for overviews of reviews and meta-research. Its structured approach promotes consistency in risk of bias assessment. However, the tool requires methodological expertise, and the absence of a quantitative scoring system can limit cross-review comparability.

JBI Critical Appraisal Tools

The Joanna Briggs Institute (JBI) Critical Appraisal Tools [9] comprise a comprehensive range of checklists designed to evaluate the methodological quality of a wide range of study designs, including randomised controlled trials, cohort studies, case-control studies, qualitative research, and case reports. Furthermore, JBI includes a dedicated critical appraisal tool for systematic reviews and research syntheses, demonstrating its utility beyond primary study designs. Each checklist contains design-specific signalling questions focusing on core domains such as clarity of inclusion criteria, validity of measurement, management of confounding and appropriateness of analysis. A key strength of the JBI framework is its adaptability, allowing consistent evaluation across heterogenous evidence. However, the tools remain largely qualitative and dependent on reviewer judgment, which introduces variability with limited capacity to assess reproducibility or workflow documentation in human-AI review settings.

MASTER scale

The MethodologicAl Standard for Epidemiological Research (MASTER) scale was developed to address the fragmentation and lack of comparability that results from using multiple design-specific quality assessment tools (e.g., separate tools for RCTs, cohort, and case-control studies)[10]. The scale provides a unified 36-item checklist categorized into seven standards like equal recruitment, equal retention, equal ascertainment, equal implementation, equal prognosis, sufficient analysis, and temporal precedence that assesses methodological quality across a broad spectrum of analytic designs. A key feature of MASTER scale is that it yields a quantitative, continuous score for quality, enabling direct cross-study comparability a function actively discouraged by tools like AMSTAR 2. By offering a comprehensive, continuous scoring system, the MASTER scale is particularly valuable in meta-epidemiological research where statistical comparisons of quality across heterogenous study designs are required.

Other Frameworks: CASP, MECIR, QUOROM, and GRADE Integration

Several additional frameworks contribute to the appraisal and reporting standards of systematic reviews. The Critical Appraisal Skills Programme (CASP) provides accessible checklists that guide users through the appraisal of different study designs, emphasising clarity, critical thinking, and educational value. The Methodological Expectations of Cochrane Intervention Reviews (MECIR) standards offer detailed methodological and reporting requirements for Cochrane reviews, promoting uniformity, transparency and rigor within that collaboration. Earlier initiatives such as QUOROM (Quality of Reporting of Meta-analyses) [11] laid the foundation for the development of PRISMA, establishing structured reporting guidelines that have since become the global standard. In contrast, the GRADE (Grading of Recommendations, Assessment, Development and Evaluation) approach [12] focuses not on the review process itself but on evaluating the certainty and strength of evidence derived from it. Collectively, these frameworks enhance methodological clarity and interpretability; however, they primarily target reporting and evidence grading.

Critical Comparison

Across existing appraisal frameworks, several persistent gaps limit their ability to fully capture the quality of systematic review conduct. While tools such as PRISMA, AMSTAR 2, ROBIS, and JBI have advanced methodological and reporting standards, their scope remains primarily confined to evaluating what was reported rather than how the review was executed. Most rely on qualitative judgments based on reviewer interpretation, which introduces subjectivity and reduces inter-rater reliability. While each framework advanced systematic review methodology within its scope, none directly measures workflow quality, reproducibility, or automation transparency - factors increasingly relevant in hybrid human–AI review settings. These constraints underscore the need for appraisal tools capable of systematically evaluating the integrity and methodological fidelity of both human-led and AI-assisted systematic reviews.

The High Yield Med Quality Evaluation Tool

Conceptual Foundation: Rationale Behind Developing a Stepwise, Workflow-Based Scoring System

The High Yield Med Quality Evaluation Tool (HYMQET) is conceptually grounded in the belief that the quality of a SR is determined by the methodological integrity maintained at every sequential step of the evidence synthesis process. Unlike tools that provide only a singular, retrospective quality assessment, the HYMQET adopts a novel workflow-based, stepwise scoring system. This design is intentionally structured to move beyond a simple pass/fail grade, allowing reviewers to precisely pinpoint where methodological breakdowns occur in the SR process. By breaking the review into distinct stages, the tool provides actionable, granular feedback that is crucial for both training human reviewers and benchmarking the fidelity and validity of automated (AI-led) systematic review methods against established human standards. This focus on process rigor ensures that the evaluation aligns with major international reporting standards, such as those set by the PRISMA statement.

Core Domains and Scoring Structure

The HYMQET assesses systematic review quality across five core domains, which collectively span the entire SR lifecycle from inception to publication. These domains are Query Development, Screening Quality, Field Selection for Data Extraction, Full-Text Data Extraction, and Manuscript Writing. The scoring mechanism employs a 5-point Likert scale (where 1 = Poor and 5 = Excellent) for each domain, enabling a quantitative assessment. Crucially, each score level is defined by explicit qualitative criteria that detail the required methodological components. For example, a score of 5 in the Query Development domain requires a ”fully optimized query with precise operators, MeSH terms, and filters,” ensuring highly relevant results, whereas a low score reflects a query lacking essential terms. In addition to the numerical score, the evaluation protocol mandates that raters provide free-text global comments on strengths and weaknesses. This hybrid approach ensures that the quantitative scores are supported by rich, qualitative justification, significantly enhancing the depth of the quality appraisal. A detailed description is found in supplementary material.

Example Application: Comparing Human vs AI-Led Systematic Review Workflows

The primary application of the HYMQET was demonstrated in a two-arm comparative study designed to assess its utility in benchmarking automated systematic review workflows [13]. The HYMQET facilitated a rigorous, head-to-head comparison between a conventional human-led pipeline and an AI-Powered Research Automation (AIPRA) pipeline, with both teams independently addressing the research question: ”What is the role of large language models in glaucoma diagnosis?” Following the established HYMQET protocol, three independent, blinded domain experts assessed the final manuscripts, data extraction tables, and included full-text articles from both workflows across the five core domains. The HYMQET scores established the human-led process as the gold standard (mean score 74.4%) against which the AIPRA output was benchmarked (mean score 65.3%). Critically, this application proved that AIPRA was non-inferior in overall quality (mean difference -9.3%). Furthermore, the study highlighted the tool's ability to capture efficiency; AIPRA completed the entire SR in approximately two hours compared to one month for the human team (a 375 × speed increase). This dual focus on quality and efficiency establishes the HYMQET as a vital instrument for the external validation and quality control of emerging automated tools in evidence synthesis.

Reliability and Reproducibility

The development of the HYMQET incorporates strong measures to ensure its reliability and reproducibility, which are essential for a robust evaluation tool. The use of multiple, independent, and blinded domain experts serves to determine inter-rater reliability, a critical metric for any quality assessment instrument. Furthermore, the detailed, pre-defined qualitative criteria corresponding to each point on the 5-point Likert scale are instrumental in standardizing the assessment process. By reducing ambiguity and subjective interpretation in scoring, these criteria reinforce the objectivity of the tool, enabling its consistent application. This structured, step-by-step approach positions the HYMQET as a highly scalable framework suitable for reliably benchmarking a variety of systematic review methodologies going forward. Nevertheless, the current version of the tool does not yet address certain methodological domains central to systematic review rigor, such as protocol registration, bias evaluation, and heterogeneity handling. Recognizing these gaps highlights areas for refinement to ensure a more comprehensive and integrative assessment framework in future versions.

Comparative Perspective: How It Differs from Existing Tools

The HYMQET fundamentally diverges from established frameworks by shifting the focus of quality appraisal from the reporting of outcomes to the integrity of the review generation process (Supplementary Table). This difference is critical for maintaining methodological rigor in the era of automated evidence synthesis. The HYMQET's quantitative, step-wise nature allows reviewers to precisely pinpoint methodological breakdowns (e.g., scoring a '2' on Query Development while scoring a '5' on Manuscript Writing). This contrasts sharply with a tool like AMSTAR 2, which may render an entire review's results "Critically Low Confidence" based on a single methodological flaw, without providing granular insight into the process's strengths. By focusing on the underlying fidelity of the workflow, the HYMQET offers a verifiable and transparent metric for the external validation and quality control of emerging automated systematic review methods.

Conclusion

The HYMQET successfully bridges a critical gap in systematic review appraisal by shifting the evaluation focus from the final reported output to the integrity and fidelity of the methodological process. Existing appraisal frameworks, while foundational for reporting and retrospective quality assessment, lack the structure to quantitatively measure the step-wise rigor and reproducibility of a review, especially within hybrid human-AI environments. The HYMQET's innovative use of a quantitative, workflow-based scoring system provides actionable, granular feedback at five critical stages of evidence synthesis. This structure makes it uniquely suited not only for training human reviewers but also for the external validation and quality control of automated systematic review pipelines. Ultimately, the HYMQET establishes a necessary new standard for methodological rigor. By ensuring transparent and verifiable workflows, it is an essential instrument for the future of reproducible and high-quality AI-enhanced evidence synthesis that directly informs clinical guidelines and health policy.

Conflict of Interest

The authors declare that they have no competing interests.

Acknowledgements

There are no acknowledgements.

Financial Support

There was no funding.

References

[1] ↑ Moosapour, H., F. Saeidifard, M. Aalaa, A. Soltani, and B. Larijani, The rationale behind systematic reviews in clinical medicine: a conceptual framework. J Diabetes Metab Disord, 2021. 20(1): p. 919-929.

[2] ↑ van der Braak, K., P. Heus, C. Orelio, F. Netterström-Wedin, K.A. Robinson, H. Lund, and L. Hooft, Perspectives on systematic review protocol registration: a survey amongst stakeholders in the clinical research publication process. Syst Rev, 2023. 12(1): p. 234.

[3] ↑ Kwong, J.C.C., A. Khondker, K. Lajkosz, M.B.A. McDermott, X.B. Frigola, M.D. McCradden, M. Mamdani, G.S. Kulkarni, and A.E.W. Johnson, APPRAISE-AI Tool for Quantitative Evaluation of AI Studies for Clinical Decision Support. JAMA Netw Open, 2023. 6(9): p. e2335377.

[4] ↑ Cabello, J.B., M. Torralba, M. Maldonado Fernandez, M. Ubeda, E. Ansuategui, L. Ramos-Ruperto, J. Emparanza, I. Urreta, M. Iglesias, and J. Pijoan, CRITICAL APPRAISAL TOOLS FOR ARTIFICIAL INTELLIGENCE CLINICAL STUDIES: A SCOPING REVIEW. Journal of Medical Internet Research, 2025.

[5] ↑ Page, M.J., J.E. McKenzie, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, L. Shamseer, J.M. Tetzlaff, E.A. Akl, S.E. Brennan, R. Chou, J. Glanville, J.M. Grimshaw, A. Hróbjartsson, M.M. Lalu, T. Li, E.W. Loder, E. Mayo-Wilson, S. McDonald, L.A. McGuinness, L.A. Stewart, J. Thomas, A.C. Tricco, V.A. Welch, P. Whiting, and D. Moher, The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ, 2021: p. n71.

[6] ↑ Shea, B.J., B.C. Reeves, G. Wells, M. Thuku, C. Hamel, J. Moran, D. Moher, P. Tugwell, V. Welch, E. Kristjansson, and D.A. Henry, AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ, 2017: p. j4008.

[7] ↑ Whiting, P., J. Savović, J.P. Higgins, D.M. Caldwell, B.C. Reeves, B. Shea, P. Davies, J. Kleijnen, and R. Churchill, ROBIS: A new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol, 2016. 69: p. 225-34.

[8] ↑ Shea, B.J., J.M. Grimshaw, G.A. Wells, M. Boers, N. Andersson, C. Hamel, A.C. Porter, P. Tugwell, D. Moher, and L.M. Bouter, Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol, 2007. 7: p. 10.

[9] ↑ Hilton, M., JBI critical appraisal checklist for systematic reviews and research syntheses (product review). Journal of the Canadian Health Libraries Association / Journal de l'Association des bibliothèques de la santé du Canada, 2024. 45(3).

[10] ↑ 1Stone, J.C., K. Glass, J. Clark, M. Ritskes-Hoitinga, Z. Munn, P. Tugwell, and S.A.R. Doi, The MethodologicAl STandards for Epidemiological Research (MASTER) scale demonstrated a unified framework for bias assessment. J Clin Epidemiol, 2021. 134: p. 52-64.

[11] ↑ Moher, D., D.J. Cook, S. Eastwood, I. Olkin, D. Rennie, and D.F. Stroup, Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of Reporting of Meta-analyses. Lancet, 1999. 354(9193): p. 1896-900.

[12] ↑ Guyatt, G.H., A.D. Oxman, G.E. Vist, R. Kunz, Y. Falck-Ytter, P. Alonso-Coello, and H.J. Schünemann, GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. Bmj, 2008. 336(7650): p. 924-6.

[13] ↑ Musleh, A., N. Alwisi, H. Abu Serhan, A. Toubasi, L. Malkawi, and S.A. Alryalat, Artificial Intelligence Powered Research Automation (AIPRA) Versus Human Expert: A Two-Arm Ophthalmology Comparative Study. medRxiv, 2025: p. 2025.10.27.25338904.