A new peer reviewer? Comparing AI with human performance in randomized controlled trial risk-of-bias assessment

Jonathan Lettner[1,2,B–F], Marko Ostojic[3,B,C,E,F], Aleksandra Królikowska[4,5,A,B,E,F], Mahmut Enes Kayaalp[1,6,E,F], Nikolai Ramadanov[1,2,E,F], Robert Prill[1,2,A,C,E,F]

[1] Faculty of Health Sciences Brandenburg, Brandenburg Medical School Theodor Fontane, Brandenburg an der Havel, Germany
[2] Center of Orthopaedics and Traumatology, University Hospital Brandenburg/Havel, Brandenburg Medical School Theodor Fontane, Brandenburg an der Havel, Germany
[3] Osteon Orthopedics and Sports Medicine Clinic, Mostar, Bosnia and Herzegovina
[4] Physiotherapy Research Laboratory, University Centre of Physiotherapy and Rehabilitation, Faculty of Physiotherapy, Wroclaw Medical University, Poland
[5] Joanna Briggs Institute (JBI) Centre for Evidence-Based Healthcare Poland, University of Adelaide, Australia
[6] Department of Orthopedics and Traumatology, University of Health Sciences, Istanbul Fatih Sultan Mehmet Training and Research Hospital, Turkey

A – research concept and design; B – collection and/or assembly of data; C – data analysis and interpretation;
D – writing the article; E – critical revision of the article; F – final approval of the article

Advances in Clinical and Experimental Medicine, ISSN 1899–5276 (print), ISSN 2451–2680 (online) *Adv Clin Exp Med*. 2026

Address for correspondence
Robert Prill
E-mail: Robert.Prill@mhb-fontane.de

Funding sources
None declared

Conflict of interest
None declared

Received on August 24, 2025
Reviewed on December 18, 2025
Accepted on December 23, 2025

Published online on April 9, 2026

Cite as
Lettner J, Ostojic M, Królikowska A, Kayaalp ME, Ramadanov N, Prill R. A new peer reviewer? Comparing AI with human performance in randomized controlled trial (RCT) risk-of-bias assessment [published online as ahead of print on April 9, 2026]. *Adv Clin Exp Med*. 2026. doi:10.17219/acem/216070

DOI
10.17219/acem/216070

Abstract

Background. Risk-of-bias (RoB) assessment is essential for evidence synthesis but remains time-consuming and inherently subjective. Artificial intelligence (AI) may improve the efficiency of systematic reviews; however, its reliability in reproducing expert RoB judgements remains uncertain.

Objectives. To compare the performance of AI models and human raters in RoB assessment of randomized controlled trials (RCTs) using the revised Joanna Briggs Institute (JBI) critical appraisal tool.

Materials and methods. Thirteen RCTs published between 2023 and 2025 in orthopedic journals were independently assessed by 2 human raters (an expert (R1) and a novice (R2)) and 2 AI models (ChatGPT-4.0 (CGPT) and DeepSeek-R1 (DS)) using the 13-domain JBI checklist. Deep-reasoning functionalities (e.g., chain-of-thought prompting) were applied. Inter-rater agreement, deviations from the expert assessment (reference standard), and binary disagreements (e.g., Yes vs No) were analyzed to evaluate consistency.

Results. The AI models demonstrated high inter-model agreement (91%), exceeding human–AI agreement (CGPT vs R1: 64%; DS vs R1: 68%). However, both AI systems showed substantial divergence from expert judgements in interpretive domains, including allocation concealment (Q2), blinding (Q7), and overall trial design (Q13), with deviation rates ranging from 30% to 38.5%. Binary decision reversals were more frequent in AI assessments (CGPT: 8.9%; DS: 7.7%) than in the human comparison (R2 vs R1: 2.4%). Human raters showed stronger agreement in contextual interpretation (R1–R2: 89.3%), whereas AI models performed better in rule-based domains (Q8/Q9: 100% agreement).

Conclusions. AI can reliably support the automation of objective components of RoB assessment but remains limited in handling interpretive, context-dependent judgements. A hybrid approach combining AI-assisted pre-screening with expert evaluation may enhance the scalability of systematic reviews without compromising methodological rigor.

Key words: systematic review, randomized controlled trials, critical appraisal, artificial intelligence, risk of bias

Highlights

- Artificial intelligence (AI) can support risk-of-bias assessment in systematic reviews.
- AI models show high inter-model agreement but moderate agreement with human reviewers.
- AI performs well in rule-based domains but struggles with subjective judgements.
- Hybrid AI–human approaches may improve systematic review efficiency and quality.

Background

The integrity of scientific research depends on the rigorous appraisal of methodological quality and the minimization of bias.[1,2] Randomized controlled trials (RCTs) are considered the gold standard for evaluating interventions across clinical and experimental settings. However, even well-designed RCTs may be affected by subtle sources of systematic error, including selection bias, performance bias, detection bias, and reporting bias, which can compromise internal validity and the reliability of subsequent meta-analyses.[3] In peer review and evidence synthesis, critical appraisal tools such as the revised Joanna Briggs Institute (JBI) checklist are indispensable for identifying these threats and determining the overall risk of bias (RoB) in trial reports.[4] By applying standardized criteria for randomization, allocation concealment, blinding, outcome measurement, and statistical analysis, RoB assessment not only safeguards the validity of individual studies but also underpins the credibility of clinical guidelines.[5]

Concurrently, artificial intelligence (AI) has emerged as a transformative force across multiple domains of scientific research. Machine learning (ML) and large language models (LLMs) have demonstrated substantial capabilities in natural language understanding, data extraction, and the interpretation of complex experimental designs.[6] In the context of systematic reviews, AI-assisted screening and data extraction have been shown to accelerate study identification, reduce reviewer workload, and, in some cases, improve the sensitivity of trial detection.[7] Despite these advances, the application of AI to formal RoB assessment remains underexplored. A key question is whether an algorithmic assessor can reliably replicate – or at least meaningfully complement – the judgements of experienced human reviewers when evaluating the methodological rigor of RCTs.[8]

Objectives

Integrating the established rigor of human critical appraisal with the efficiency and scalability of AI may enhance both the speed and consistency of RoB assessment. This study aimed to compare the performance of AI models and human reviewers in evaluating RoB in RCTs using the revised JBI critical appraisal tool. Specifically, we sought to determine the extent of agreement between AI- and human-generated assessments, identify domains in which AI aligns with or diverges from expert judgement, and evaluate the potential role of AI in supporting evidence synthesis workflows while preserving methodological rigor.

Materials and methods

Study design

This retrospective, comparative methodological analysis examined 13 RCTs published in *Knee Surgery, Sports Traumatology, Arthroscopy* (n = 10)[9–18] and the *Journal of Experimental Orthopaedics* (n = 3).[19–21] The objective was to systematically compare RoB assessments conducted by human reviewers and AI-based models. As far as possible, the study adhered to the STARD (Standards for Reporting of Diagnostic Accuracy Studies) reporting guidelines to ensure transparent and standardized reporting of methods and results.[22]

Study selection

The included studies were published between January 2023 and March 2025, comprising 6 studies from 2023 ([9,11,16–19]), 4 from 2024 ([10,12,14,21]), and 3 from 2025 ([13,15,20]). Inclusion criteria were as follows: original RCTs, full-text availability in English, and clear reporting of randomization and blinding procedures. Exclusion criteria included observational studies, review articles, and studies lacking sufficient methodological detail.

Reviewers

Two independent human reviewers conducted the assessments: 1 experienced reviewer (R1) with more than 60 completed peer reviews in orthopedic research, and 1 less experienced reviewer (R2) with fewer than 20 completed peer reviews.

AI-based risk assessments

Thirteen full-text articles were uploaded as *pdf files to ChatGPT (CGPT; OpenAI, v. 4, April 2025 release) and DeepSeek (DS; Standard R1 2025 version). No plain text conversions or copy-paste operations were performed. In both

Table 1. Software specifications

Aspect	ChatGPT (OpenAI, GPT-4, April 2025)	DeepSeek (Standard R1 2025)
Interface	OpenAI REST API v1.0	HTTPS REST endpoint with JSON responses
Implementation	Python 3.10 (CPython) client using requests v2.31.0 and built-in JSON parser	Proprietary 12-billion-parameter Transformer backbone
Document upload	Binary multipart upload via OpenAI Files API	Multipart/form-data upload
Prompt configuration	System prompt enabling chain-of-thought and deep-reasoning capabilities by default	Prompt flag to deactivate all DeepThink modules
Limits and output	128 000-token limit per request; articles ranged from 5,000 to 12,000 tokens	JSON response objects for each JBI domain, including domain, rating, and confidence score

systems, all deep reasoning functionalities (i.e., Chain-of-Thought in CGPT and DeepThink in DS) were enabled. The specific prompt issued to each model was: "Please assess the attached manuscript using the revised JBI critical appraisal tool for RoB in randomized controlled trials."

Each model subsequently provided domain-level evaluations for all 13 items of the JBI Critical Appraisal Tool for RCTs (Revised Checklist 2023). The specific software is presented in Table 1.

Risk of bias assessment

The RoB in the 13 RCTs was assessed using the revised JBI Critical Appraisal Checklist for RCTs,[4] which covers 13 domains central to methodological rigor and internal validity. Key aspects include the adequacy of randomization and concealment of group allocation, as well as baseline comparability between treatment groups. The presence and implementation of blinding were examined for participants, treatment providers, and outcome assessors.

Further evaluation focused on whether treatment groups were managed identically apart from the intervention of interest, and whether outcomes were measured consistently and with reliable instruments. Completeness of follow-up and handling of attrition were considered, including whether group differences in follow-up were appropriately described and analyzed.

Additional domains addressed the analysis of participants in their originally assigned groups, the appropriateness of statistical methods, and the adequacy of the overall trial design, including the management of deviations from standard RCT frameworks (e.g., individual randomization or parallel-group designs). Each domain was classified as Yes, Unclear/Not applicable, or No.

Data extraction and management

All assessments (study ID, journal, year, reviewer or AI model, domain, and RoB classification) were recorded in a standardized spreadsheet (Microsoft Excel for Microsoft 365, v. 2405; Microsoft Corp., Redmond, USA). To ensure blinding, author names, journal titles, and publication details were anonymized prior to assessment. The results were then submitted to an independent third party, who compiled the data into the final dataset. Reviewers had no access to each other's evaluations or to those generated by the AI models.

Statistical analyses

Analyses were descriptive. For each JBI domain across the 13 trials, inter-assessor agreement was expressed as the percentage of identical ratings across all 4 raters and, separately, as percentage agreement relative to the expert reviewer (R1).

Overall deviation from R1 was calculated for each comparator (R2, CGPT, DS) across the full rating matrix (13 domains × 13 trials = 169 observations). Ratings were numerically coded (Yes = 0; Unclear/Not applicable = 1; No = 2) to enable the calculation of pairwise inter-rater correlations.

Both Pearson's correlation coefficients (r) and Kendall's tau were computed to quantify the strength of association between raters. In addition, "binary flips" were identified to characterize extreme discordance, defined as a direct shift between 0 (Yes) and 2 (No) for the same item. All analyses, including percentage agreement, deviation rates, and correlation coefficients, were performed using Microsoft Excel for Microsoft 365, v. 2405 (Microsoft Corp.).

Results

Overall inter-rater agreement

Across the 13 RCTs, the 4 assessors (R1, R2, CGPT, and DS) demonstrated variable levels of agreement across the 13 JBI checklist domains. Table 2 summarizes the percentage agreement among all 4 reviewers for each domain. Perfect agreement (100%) was observed for Q8 ("Were outcomes measured in the same way for treatment groups?") and Q9 ("Were outcomes measured reliably?"). High agreement (>84%) was noted for Q1 (randomization; 84.6%), Q6 (identical treatment apart from the intervention; 92.3%), Q10 (follow-up completeness; 84.6%), and Q12 (appropriateness of statistical analysis; 84.6%). In contrast, the lowest levels of agreement were observed for Q13 (trial design adequacy; 38.5%), Q7 (blinding of outcome assessors; 46.2%), and Q3 (baseline comparability; 53.8%).

Table 2. Inter-reviewer agreement

Domain	Agreed assessment	Total assessment	Agreement (%)
Q1	11	13	84.60
Q2	8	13	61.50
Q3	7	13	53.80
Q4	9	13	69.20
Q5	9	13	69.20
Q6	12	13	92.30
Q7	6	13	46.15
Q8	13	13	100
Q9	13	13	100
Q10	11	13	84.60
Q11	9	13	69.20
Q12	11	13	84.60
Q13	5	13	38.46

Deviation from expert reviewer

Table 3 presents the deviations of CGPT, DS, and R2 relative to R1, which was treated as the reference standard. Both CGPT and DS showed the greatest deviations from R1 in domains Q2, Q3, Q7, and Q13, with deviation rates of 38.5%. The highest level of agreement between CGPT and R1 was observed for Q8, Q9, and Q11, with no deviations (0%), indicating complete concordance. ChatGPT also demonstrated strong agreement in Q6 and Q10, each with a deviation rate of 7.7%. Overall, the mean deviation between CGPT and R1 was 17.8%.

DeepSeek demonstrated a deviation pattern similar to that of CGPT. The greatest discrepancies relative to R1 were observed in domains Q2, Q3, and Q7 (each 38.5%), as well as in Q13 (30.8%). In contrast, DS achieved complete agreement with R1 in Q8, Q9, Q11, and Q12. The mean deviation between DS and R1 was 17.2%.

The novice reviewer (R2) demonstrated the highest level of agreement with the expert reviewer (R1). The greatest deviations were observed in domains Q2 and Q13, both at 23.1%.

Complete agreement (0% deviation) was achieved for Q8, Q9, Q10, and Q12. Overall, the mean deviation between R2 and R1 was 10.7%.

Pairwise agreement comparison

When comparing the 2 AI systems directly, CGPT and DS agreed on 91.1% of ratings overall. Perfect concordance (100%) was observed for 6 domains (Q1, Q3, Q6, Q8, Q9, and Q11), whereas the lowest agreement occurred in Q4 (61.5%) and Q7 (76.9%). These findings indicate a high level of inter-model consistency between the 2 AI systems.

Agreement between the human reviewers (R1 and R2) was also high (89.3%), with perfect concordance observed for Q8–Q10 and Q12 (100%; Table 4). Notably, the 2 AI models showed slightly higher agreement with each other than with human reviewers.

Correlation of reviewer pairs

Kendall's tau correlations showed positive concordance among all raters, with values ranging from 0.61 to 0.79. The highest agreement was observed between CGPT and DS ($\tau = 0.787$), followed by DS and R2 ($\tau = 0.706$), DS and R1 ($\tau = 0.684$), CGPT and R2 ($\tau = 0.627$), and CGPT and R1 ($\tau = 0.612$). The strongest concordance overall was observed between R2 and R1 ($\tau = 0.792$). Pearson's correlation

Table 3. Deviation rates from R1

Domain	Cases	CGPT Dev.	DS Dev.	R2 Dev.	CGPT [%]	DS [%]	R2 [%]
Q1	13	2	2	2	15.40	15.40	15.40
Q2	13	5	5	3	38.50	38.50	23.10
Q3	13	5	5	2	38.50	38.50	15.40
Q4	13	2	3	1	15.40	23.10	7.70
Q5	13	2	2	2	15.40	15.40	15.40
Q6	13	1	1	1	7.70	7.70	7.70
Q7	13	5	5	2	38.50	38.50	15.40
Q8	13	0	0	0	0.00	0.00	0.00
Q9	13	0	0	0	0.00	0.00	0.00
Q10	13	1	2	0	7.70	15.40	0.00
Q11	13	0	0	2	0.00	0.00	15.40
Q12	13	2	0	0	15.40	0.00	0.00
Q13	13	5	4	3	38.50	30.80	23.10
Mean					17.77	17.18	10.66

Expert (R1) and novice (R2) reviewer; CGPT – ChatGPT; DS – DeepSeek.

Table 4. Pairwise agreement between human reviewers and AI models

Domain	R2 = R1 (n/13)	R2–R1 [%]	CGPT = DS (n/13)	CGPT–DS [%]
Q1	11	84.60	13	100
Q2	10	76.90	11	84.60
Q3	11	84.60	13	100
Q4	12	92.30	8	61.54
Q5	11	84.60	12	92.30
Q6	12	92.30	13	100
Q7	11	84.60	10	76.90
Q8	13	100	13	100
Q9	13	100	13	100
Q10	13	100	12	92
Q11	11	84.60	13	100
Q12	13	100	11	84.60
Q13	10	76.90	12	92.30
Overall	151/169	89.34	154/169	91.12

Expert (R1) and novice (R2) reviewer; CGPT – ChatGPT; DS – DeepSeek.

coefficients showed a similar pattern. The strongest association was observed between the 2 AI models (r = 0.91), followed by R2 and R1 (r = 0.89) and CGPT vs R2 (r = 0.67). The lowest correlation was observed between CGPT and R1 (r = 0.64) (Table 5).

Magnitude of discrepancies ("binary flips")

To assess the severity of disagreement, we examined cases of "binary flips", defined as instances in which R1 rated an item as Yes (0), while another rater assigned a rating of No (2), representing a complete reversal in judgement. Among the 169 rated items, such major discrepancies occurred only 4 times (2.4%) between the 2 human reviewers (R1 and R2). In contrast, CGPT exhibited 15 "binary flips" (8.9%) and DS 13 (7.7%) relative to R1 (Table 6). A detailed analysis of the 13 RCTs is presented in Supplementary Table 1.

Discussion

Systematic reviews are a cornerstone of evidence-based clinical decision-making, particularly in specialties such as orthopedics, traumatology, sports medicine, and rehabilitation, where treatment decisions are increasingly complex and must be guided by high-quality evidence.[23–25] At the core of every robust review lies the critical appraisal of included studies, which aims to determine the extent to which bias has been minimized through appropriate study design, conduct, and analysis.[26–28]

One of the key challenges in RoB assessment is that bias must be evaluated not only at the study level but also at the level of individual outcomes and specific results. This distinction is critical, as a single trial may apply different measurement methods or bias-mitigation strategies across outcomes, leading to varying levels of bias within the same study. To address this complexity, the JBI Effectiveness Methodology Group published a revised set of critical appraisal tools in 2023, specifically designed to facilitate multi-level RoB assessment.[29,30] The updated JBI checklist for RCTs allows separate evaluation of multiple outcomes (up to 7 per study) and enables result-level assessment (Questions 10–12) for up to 3 specific results per outcome. In principle, assessments should focus on outcomes directly relevant to the review question. However, as the present study did not pre-specify a primary outcome, a comprehensive approach – evaluating all reported outcomes across the included trials – was adopted.

Within this framework, a key limitation of algorithmic evaluation became apparent. Both DS and CGPT did not consistently apply the outcome-specific approach embedded

Table 5. Inter-rater correlation (Pearson's correlation coefficient)

CGPT–DS	CGPT–R2	CGPT–R1	DS-R2	DS-R1	R2-R1
0.91	0.67	0.64	0.76	0.68	0.89

Expert (R1) and novice (R2) reviewer; CGPT – ChatGPT; DS – DeepSeek.

Table 6. Deviation patterns from R1

Deviation R1 → R2	N	Deviation R1 → CGPT	N	Deviation R1 → DS	N
0 → 1	5	0 → 1	6	0 → 1	7
0 → 2	4	0 → 2	10	0 → 2	9
1 → 0	2	1 → 0	0	1 → 0	1
2 → 0	0	2 → 0	5	2 → 0	4
1 → 2	4	1 → 2	8	1 → 2	6
2 → 1	4	2 → 1	0	2 → 1	2

Expert (R1) and novice (R2) reviewer; CGPT – ChatGPT; DS – DeepSeek.

in the 2023 JBI methodology. Instead, they tended to default to a more general, study-level interpretation, failing to capture the level of nuance required by the revised tool. This finding likely reflects not only limited alignment with current methodological standards but also highlights a broader challenge for AI systems in handling multi-level and structurally complex appraisal frameworks.

Where the standards are solidly in place, operationalized standards – such as the use of validated outcome measures or open reporting of statistical practices – AI consistently equaled human reviewers' precision. For example, questions that are highly structured and rule-based – such as assessing randomization (Q1), treatment consistency (Q6), and outcome measurement reliability (Q8–Q12) – showed high agreement between AI and human reviewers, with CGPT and DS reaching 100% agreement on some items (e.g., Q8 and Q9). This suggests that AI can reliably perform tasks that are checklist-driven and predictable.

In contrast, domains requiring interpretive reasoning, contextual judgement, or the integration of dispersed information – such as allocation concealment (Q2), baseline comparability (Q3), blinding of outcome assessors (Q7), and the adequacy of trial design (Q13) – were associated with lower agreement between AI and human reviewers. These tasks often rely on information that is distributed across different sections of a manuscript (e.g., appendices, figure legends, or indirectly reported text), requiring synthesis of implicit or non-standardized cues. In such contexts, AI models demonstrated limited ability to consistently integrate and interpret these signals. This pattern underscores a fundamental limitation: while AI performs well in structured, rule-based assessments, it remains challenged by tasks that require inference, nuanced interpretation, and context-sensitive reasoning. Consequently, transformer-based models appear well suited to checklist-driven evaluations but less reliable in domains requiring deeper methodological judgement.[31]

Nevertheless, in situations where methodological reporting was less transparent and the language more indirect, a different pattern emerged. Even the expert reviewer (R1) occasionally encountered difficulty in determining whether baseline characteristics were truly comparable, whether allocation concealment had been adequately maintained, or whether blinding of treatment providers was sufficiently described. In many such cases, the relevant information was presented indirectly or dispersed across different sections of the manuscript (e.g., appendices, footnotes, or figure legends). AI systems were more likely to overlook or misinterpret these cues, highlighting an important limitation: while AI is well suited to rule-based pattern recognition, it remains less effective at synthesizing contextual information and interpreting implicitly reported details.

These difficulties were strongly apparent in those situations that demanded interpretive sensitivity. For instance, where blinding methods were referred to vaguely – i.e., describing an injector as blinded without explicitly asserting the same – AI systems did not typically signal this vagueness. Reviewer 1, however, could sense the vagueness and appropriately flag the RoB as unknown. This is a root failure of current AI tools: they lack the implicit inference and contextual flexibility provided by human evaluators – something that cannot be fixed merely by ramping up model scale or computational power.

In a remarkable twist, agreement between CGPT and DS was consistently high – even when both disagreed with human decision-making. This implies that differences in training data or model architecture might be less critical than structure of prompts or the application of reasoning aids. When reasoning aids like Chain-of-Thought or DeepThink were turned off, both models fell back to a strict pattern-matching style, mapping textual phrases onto fixed checklist bins. Though this reduces interpretive flexibility, it has potential for innovation: future systems could benefit from modular prompting, in which more penetrating processes of reason are accessed only for more ambiguous items.[8]

Human evaluators, in contrast, demonstrated greater consistency and precision when assessing ambiguous or context-dependent cases. Notably, agreement between the expert reviewer and the less experienced human rater was higher than that observed between either human reviewer and the AI systems. These findings suggest that even moderate methodological training may enable human reviewers to outperform current AI systems in more complex appraisal domains, particularly those requiring interpretive judgement and experience-based pattern recognition.[8]

A deeper analysis of the types of disagreements further highlights risks to reliability for AI-driven appraisal. Reversals by binary – cases where the AI made a diametrically opposite choice from the specialist (e.g., Yes vs No) – were of particular concern. ChatGPT exhibited 15 such reversals (8.9% of its choices), and DS exhibited 13 (7.7%), whereas the 2nd human evaluator had only 4 (2.4%). These are not small discrepancies – they are fundamentally different conclusions about RoB, and they are occurring with enough frequency to raise serious questions about the validity of AI-based assessments. Most of these were in studies that had non-conventional or indirect reporting strategies, highlighting the difficulty of current models with document navigation and interpretive synthesis.[32,33]

These limitations are not solely technical but also conceptual. The decision-making processes of deep learning models remain largely opaque; even when outputs appear accurate, the underlying reasoning is often not accessible. Although AI systems represent input text in high-dimensional vector spaces, they do not readily indicate which elements of the input most strongly influenced a given output. This lack of transparency limits auditability and may undermine trust in individual judgements.

Techniques such as attention visualization or relevance propagation may help improve interpretability by providing insight into how specific conclusions are generated and where potential errors arise.[34] The integration of AI into

critical appraisal workflows also raises important ethical considerations. A key concern is automation bias – the tendency of users to accept AI-generated outputs uncritically, even when they may conflict with expert judgement. This risk is particularly relevant in systematic reviews, where uncorrected errors can propagate through evidence synthesis and potentially influence clinical recommendations. To ensure credibility, transparency is essential, particularly with respect to model provenance, prompt configuration, and versioning. Furthermore, the proprietary nature of many commercial large language models poses additional challenges, including limited reproducibility, restricted accessibility in low-resource settings, and the potential to exacerbate global disparities in research capacity. Without the development of open standards for benchmarking and more transparent or community-driven approaches to model development and licensing, the benefits of AI in evidence synthesis may be unevenly distributed.[35,36] At present, there is no widely adopted framework for systematically evaluating the performance of AI in RoB assessment. Addressing this gap will be essential before such tools can be safely and reliably integrated into evidence synthesis workflows.[37]

Limitations

This study has several limitations. First, the relatively small, field-specific sample of 13 RCTs published in orthopedic journals may not fully capture the heterogeneity of reporting practices across medical disciplines. Second, the use of deep-reasoning approaches (e.g., chain-of-thought prompting) provided an experimental framework for comparing models but may not fully reflect the performance of AI systems under alternative or more advanced configurations. Third, the use of a single expert reviewer as the reference standard introduces an element of subjectivity. Although this reflects common practice in RoB assessment, future studies should incorporate multiple expert reviewers with consensus adjudication to enhance validity. In addition, in routine systematic review workflows, a third-party adjudicator is typically consulted in cases of disagreement – a step not implemented in the present study.

Fourth, the study did not evaluate practical workflow metrics, such as appraisal time, resource utilization, or user satisfaction, which limits its applicability to real-world implementation. Finally, given the rapid evolution of AI systems and the growing landscape of open-source alternatives, these findings represent a temporal snapshot. Ongoing reassessment will be necessary as model capabilities and prompting strategies continue to evolve, potentially altering the strengths and limitations of AI-assisted critical appraisal.

In addition, only a single standardized prompt was used to instruct the AI models. Even minor variations in prompt phrasing may influence model outputs. This study did not assess output reproducibility or reliability, as each model was queried only once per article, and prompts were not repeated to evaluate response consistency across multiple runs. Importantly, this design was intended to reflect a pragmatic scenario in which less experienced researchers may rely on a single prompt when using AI tools for critical appraisal tasks.

Moreover, a further limitation relates to the inability of AI systems to critically evaluate their own assumptions and outputs. In this study, the AI models were queried regarding the version of the JBI critical appraisal tool they were applying. Both models indicated that they were using the revised 2023 version and provided a link to the corresponding document, suggesting that the correct checklist was being referenced. However, this apparent self-verification does not necessarily ensure that the tool was applied accurately or in accordance with its intended methodological framework.

Conclusions

AI-enabled RoB assessment shows considerable promise for automating well-defined, checklist-based components of critical appraisal. However, human expertise remains essential for capturing methodological nuance and context-dependent judgements. A pragmatic path forward may lie in a hybrid framework, in which AI supports the efficient pre-screening of objective criteria, while more complex, interpretive assessments remain the responsibility of trained human reviewers. Such an approach has the potential to enhance efficiency without compromising the methodological rigor that underpins trustworthy evidence synthesis.

Supplementary data

The supplementary materials are available at https://doi.org/10.5281/zenodo.18759418. The package contains the following files:

Supplementary Table 1. Detailed analysis of the magnitude of discrepancies.

Data Availability Statement

The datasets supporting the findings of the current study are openly available in Figshare at https://doi.org/10.6084/m9.figshare.31382911.

Consent for publication of personal information

Not applicable.

Use of AI

ChatGPT and DeepSeek were used for the risk-of-bias assessment of the 13 included RCTs.

ORCID iDs

Jonathan Lettner https://orcid.org/0009-0008-9969-6189
Marko Ostojic https://orcid.org/0000-0002-0108-5750
Aleksandra Królikowska https://orcid.org/0000-0002-6283-5500
Mahmut Enes Kayaalp https://orcid.org/0000-0002-9545-7454
Nikolai Ramadanov https://orcid.org/0000-0003-4669-8187
Robert Prill https://orcid.org/0000-0002-4916-1206

References

1. Prill R, Królikowska A, De Girolamo L, Becker R, Karlsson J. Checklists, risk of bias tools, and reporting guidelines for research in orthopedics, sports medicine, and rehabilitation. *Knee Surg Sports Traumatol Arthrosc*. 2023;31(8):3029–3033. doi:10.1007/s00167-023-07442-8
2. Martin RK, Ley C, Pareek A, Groll A, Tischer T, Seil R. Artificial intelligence and machine learning: An introduction for orthopaedic surgeons. *Knee Surg Sports Traumatol Arthrosc*. 2022;30(2):361–364. doi:10.1007/s00167-021-06741-2
3. Królikowska A, Urban N, Lech M, et al. Mapping the reporting practices in recent randomised controlled trials published in *Knee Surgery, Sports Traumatology, Arthroscopy*: A scoping review of methodological quality. *J Exp Orthop*. 2025;12(1):e70117. doi:10.1002/jeo2.70117
4. Barker TH, Habibi N, Aromataris E, et al. The revised JBI critical appraisal tool for the assessment of risk of bias for quasi-experimental studies. *JBI Evid Synth*. 2024;22(3):378–388. doi:10.11124/JBIES-23-00268
5. Prill R, Pieper D, Klugar M, Ayeni OR, Karlsson J, Lund H. Evidence-based research in orthopaedics, sports medicine and rehabilitation: Why new studies should rely on earlier work. *Knee Surg Sports Traumatol Arthrosc*. 2024;32(2):203–205. doi:10.1002/ksa.12047
6. Kayaalp ME, Ollivier M, Winkler PW, et al. Embrace responsible ChatGPT usage to overcome language barriers in academic writing. *Knee Surg Sports Traumatol Arthrosc*. 2024;32(1):5–9. doi:10.1002/ksa.12014
7. Ko S, Pareek A, Ro DH, et al. Artificial intelligence in orthopedics: Three strategies for deep learning with orthopedic specific imaging. *Knee Surg Sports Traumatol Arthrosc*. 2022;30(3):758–761. doi:10.1007/s00167-021-06838-8
8. Hughes JD, Cristiani R, Hirschmann MT, Musahl V, Eriksson K, Karlsson J. Tips and tricks for how to become a good reviewer. *Knee*. 2023;31(11):4631–4636. doi:10.1007/s00167-023-07595-6
9. Nam HS, Yoo HJ, Ho JPY, Kim YB, Lee YS. Preoperative education on realistic expectations improves the satisfaction of patients with central sensitization after total knee arthroplasty: A randomized-controlled trial. *Knee Surg Sports Traumatol Arthrosc*. 2023;31(11):4705–4715. doi:10.1007/s00167-023-07487-9
10. Ettinger M, Tuecking L, Savov P, Windhagen H. Higher satisfaction and function scores in restricted kinematic alignment versus mechanical alignment with medial pivot design total knee arthroplasty: A prospective randomised controlled trial. *Knee Surg Sports Traumatol Arthrosc*. 2024;32(5):1275–1286. doi:10.1002/ksa.12143
11. Sørensen OG, Faunø P, Konradsen L, et al. Combined anterior cruciate ligament revision with reconstruction of the antero-lateral ligament does not improve outcome at 2-year follow-up compared to isolated acl revision: A randomized controlled trial. *Knee Surg Sports Traumatol Arthrosc*. 2023;31(11):5077–5086. doi:10.1007/s00167-023-07558-x
12. Sevinc C, Gürler V, Harput G, Ocguder A, Ergen FB, Tunay VB. Blood flow restriction training with cross education for quadriceps muscle recovery after anterior cruciate ligament reconstruction: A prospective, randomized, controlled, single-blind clinical trial. *Knee Surg Sports Traumatol Arthrosc*. 2025;33(9):3088–3097. doi:10.1002/ksa.12553
13. Theeuwen DMJ, Dorling IM, Most J, et al. Patient-specific instrumentation improved clinical outcome and implant survival but is not superior compared to conventional total knee arthroplasty: Ten years follow-up of a multicenter double-blind randomized controlled trial. *Knee Surg Sports Traumatol Arthrosc*. 2025;33(4):1371–1377. doi:10.1002/ksa.12505
14. Nakamura E, Okamoto N, Masuda T, et al. Medial-pivot design does not provide superior clinical results compared to posterior-stabilized total knee arthroplasty despite kinematic differences during step-up and lunge activities: A prospective randomized controlled trial under medial tight soft tissue balance. *Knee Surg Sports Traumatol Arthrosc*. 2024;32(12):3289–3298. doi:10.1002/ksa.12399
15. Farhan-Alanie OM, Doonan J, Rowe PJ, et al. Prospective, randomised controlled trial comparing robotic arm-assisted bi-unicompartmental knee arthroplasty to total knee arthroplasty. *Knee Surg Sports Traumatol Arthrosc*. 2025;33(7):2571–2580. doi:10.1002/ksa.12644
16. Marinova M, Sundaram A, Holtham K, et al. The role of a cryocompression device following total knee arthroplasty to assist in recovery: A randomised controlled trial. *Knee Surg Sports Traumatol Arthrosc*. 2023;31(10):4422–4429. doi:10.1007/s00167-023-07455-3
17. Bollars P, Janssen D, De Weerdt W, et al. Improved accuracy of implant placement with an imageless handheld robotic system compared to conventional instrumentation in patients undergoing total knee arthroplasty: A prospective randomized controlled trial using CT-based assessment of radiological outcomes. *Knee Surg Sports Traumatol Arthrosc*. 2023;31(12):5446–5452. doi:10.1007/s00167-023-07590-x
18. Demirci H, Van Der Storm SL, Huizing NJ, et al. Watching a movie or listening to music is effective in managing perioperative anxiety and pain: A randomised controlled trial. *Knee Surg Sports Traumatol Arthrosc*. 2023;31(12):6069–6079. doi:10.1007/s00167-023-07629-z
19. Barroso Rosa S, Wilkinson M, McEwen P, et al. Skin sensory alteration and kneeling ability following cruciate retaining total knee arthroplasty are not affected by the incision position: A randomised controlled trial of simultaneous bilateral surgery. *J Exp Orthop*. 2023;10(1):145. doi:10.1186/s40634-023-00695-9
20. Yasui Y, Miyamoto W, Sasahara J, et al. No significant impact of platelet-rich plasma on recovery after Achilles tendon surgery: A double-blind randomized controlled trial. *J Exp Orthop*. 2025;12(1):e70168. doi:10.1002/jeo2.70168
21. Macedo F, Lucas J, Cunha P, et al. No difference in patient-reported outcomes or range of motion between ultracongruent and posterior stabilized total knee arthroplasty: A randomized controlled trial. *J Exp Orthop*. 2024;11(4):e70043. doi:10.1002/jeo2.70043
22. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. *BMJ*. 2015;351:h5527. doi:10.1136/bmj.h5527
23. Królikowska A, Prill R, Klugar M. Evidence-based healthcare: Bridging the gap between research and practice. *Adv Clin Exp Med*. 2025;34(2):139–147. doi:10.17219/acem/201184
24. Prill R, Królikowska A, Enes Kayaalp M, Ramadanov N, Karlsson J, Hirschmann MT. Enhancing research methods: The role of systematic and scoping reviews in orthopaedics, sports medicine and rehabilitation. *J Exp Orthop*. 2024;11(4):e70069. doi:10.1002/jeo2.70069
25. Prill R, Mouton C, Klugorová J, Królikowska A, Karlsson J, Becker R. Implementation of evidence-based medicine in everyday clinical practice. *Knee Surg Sports Traumatol Arthrosc*. 2023;31(8):3034–3036. doi:10.1007/s00167-023-07468-y
26. Munn Z, Barker TH, Moola S, et al. Methodological quality of case series studies: An introduction to the JBI critical appraisal tool. *JBI Evid Synth*. 2020;18(10):2127–2133. doi:10.11124/JBISRIR-D-19-00099
27. Porritt K, Gomersall J, Lockwood C. JBI's Systematic Reviews: Study selection and critical appraisal. *Am J Nurs*. 2014;114(6):47–52. doi:10.1097/01.NAJ.0000450430.97383.64
28. Sterne JAC, Savović J, Page MJ, et al. RoB 2: A revised tool for assessing risk of bias in randomised trials. *BMJ*. 2019;366:l4898. doi:10.1136/bmj.l4898
29. Barker TH, Stone JC, Sears K, et al. Revising the JBI quantitative critical appraisal tools to improve their applicability: An overview of methods and the development process. *JBI Evid Synth*. 2023;21(3):478–493. doi:10.11124/JBIES-22-00125
30. Barker TH, Stone JC, Sears K, et al. The revised JBI critical appraisal tool for the assessment of risk of bias for randomized controlled trials. *JBI Evid Synth*. 2023;21(3):494–506. doi:10.11124/JBIES-22-00430
31. Beck S, Kuhner M, Haar M, Daubmann A, Semmann M, Kluge S. Evaluating the accuracy and reliability of AI chatbots in disseminating the content of current resuscitation guidelines: A comparative analysis between the ERC 2021 guidelines and both ChatGPTs 3.5 and 4. *Scand J Trauma Resusc Emerg Med*. 2024;32(1):95. doi:10.1186/s13049-024-01266-2
32. Oettl FC, Pareek A, Winkler PW, et al. A practical guide to the implementation of AI in orthopaedic research, Part 6: How to evaluate the performance of AI research? *J Exp Orthop*. 2024;11(3):e12039. doi:10.1002/jeo2.12039

33. Van Dis EAM, Bollen J, Zuidema W, Van Rooij R, Bockting CL. ChatGPT: Five priorities for research. *Nature*. 2023;614(7947):224–226. doi:10.1038/d41586-023-00288-7
34. Giorgetti C, Giorgetti A, Boscolo-Berto R. Establishing new boundaries for medical liability: The role of AI as a decision-maker. *Adv Clin Exp Med*. 2025;34(10):1601–1606. doi:10.17219/acem/208596
35. Cretu C. How does ChatGPT actually work? An ML engineer explains. San Francisco, USA: Scalable Path; 2025. https://www.scalablepath.com/machine-learning/chatgpt-architecture-explained. Accessed May 16, 2025.
36. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: An overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell*. 2023;6:1169595. doi:10.3389/frai.2023.1169595
37. Dahmen J, Kayaalp ME, Ollivier M, et al. Artificial intelligence bot ChatGPT in medical research: The potential game changer as a double-edged sword. *Knee Surg Sports Traumatol Arthrosc*. 2023;31(4):1187–1189. doi:10.1007/s00167-023-07355-6