eISSN 2444-7986
DOI: https://doi.org/10.14201/orl.33188

SUPPLEMENTARY DOCUMENTATION

ARTIFICIAL INTELLIGENCE IN PEER REVIEW: ACCELERATION WITHOUT LOSING SCIENTIFIC INTEGRITY? SUPPLEMENTARY DOCUMENTATION

José Luis PARDAL-REFOYO

Director of Revista ORL. University Hospital of Salamanca. Otorhinolaryngology and Head and Neck Surgery. University of Salamanca. Department of Surgery. Faculty of Medicine. Otorhinolaryngology Area. IBSAL. Salamanca. Spain.

Correspondence: jlpardal@usal.es

Published Date: April 27, 2025
Date of publication of the fascicle: June 27, 2025

Conflict of interest: The authors declare no conflicts of interest
Images: The authors declare that they obtained the images with the permission of the patients
Rights and self-archiving policy: self-archiving of the post-print version (SHERPA/RoMEO) is allowed
CC BY-NC-ND license. Creative Commons Attribution-NonCommercial-NoDerivative 4.0 International License
University of Salamanca. Its marketing is subject to the permission of the publisher

SUMMARY: The literature confirms that artificial intelligence is increasingly integrated across all peer review stages in science and medicine—improving efficiency in screening, reviewer assignment, and reporting—yet also introduces new ethical, legal, and quality challenges that require transparent guidelines, human oversight, and ongoing policy development [1-10].

INTRODUCTION [1-10]

The peer review process is central to maintaining standards and integrity in academic publishing, particularly in the sciences and medicine. Traditionally characterized by human-led evaluation—often involving multiple stages such as initial screening, reviewer assignment, quality assessment, and decision-making—peer review has faced persistent challenges, including reviewer shortages, bias, variable quality, and increasing submission volumes. In recent years, artificial intelligence (AI) has emerged as a transformative force, offering new tools and approaches to streamline and potentially enhance various stages of peer review.

A systematic review of recent literature assisted by UnderMind reveals a broad and nuanced discussion on the integration of AI into peer review workflows. Several papers provide comprehensive overviews of how AI is being incorporated across multiple stages of the review process. Mollaki 2024 [1], details the current landscape, highlighting the absence of clear policies regarding AI use by reviewers, and summarizes both the strengths (e.g., speed, consistency) and limitations (e.g., superficial feedback, fabricated references, algorithmic bias) of AI-assisted review. Eger et al. [4] present a survey on AI-supported peer review, describing advances in automated review generation, aggregation, and meta-reviews, while identifying significant research gaps—particularly a lack of domain diversity and limited evaluation of real-world deployment and impacts on author behavior.

Doskaliuk et al. 2025 [5] evaluate the efficiency gains and ethical considerations associated with integrating AI in peer review, emphasizing the necessity of human oversight and clear guidelines to safeguard scientific integrity. Additional works, such as those by Cárdenas 2023 [3] and Carobene et al. 2023 [7], further contextualize AI's dual role in manuscript drafting and review, examining both technical potential and risks such as deskilling and fairness concerns. These overviews are complemented by empirical studies—Checco et al. 2021 [9] and Farber 2024 [10], for instance, offer evaluations of AI prototypes for manuscript screening and reviewer assignment, reporting improvements in efficiency and insights into bias but also emphasizing the enduring need for human–machine complementarity.

A notable theme across the literature is the ongoing tension between efficiency gains and ethical or practical risks. Comparative analyses highlight both the advantages (e.g., fast turnaround, error detection, expanded reviewer pools) and the inherent limitations of AI, especially in assessing novelty or complex argumentation [1, 2, 4-7, 9, 10]. Finally, several papers review evolving institutional and international policies on AI in peer review, noting that while bodies such as COPE and ICMJE have focused on authors' use of AI, guidance on reviewer-side applications remains limited and fragmented [1-3].

In summary, the current literature documents rapid advances in AI-assisted peer review across all workflow stages, while consistently urging transparent guidelines, continual human oversight, and further research on the downstream impacts of these technologies on scholarly publishing.

KEY FINDINGS BY PROCESS STAGE

1. Multi-Stage AI Integration in Peer Review

Screening/Triage: AI tools perform formatting, plagiarism, and topicality checks, speeding up initial manuscript triage and filtering out-of-scope submissions [1, 3, 4, 6, 7, 9].

Reviewer Assignment: AI-driven matching with embeddings and coauthorship/network data shows 42 % overlap with human editors and expands reviewer pool, but accuracy varies by field and concerns about interpretability and bias remain [1, 4, 9, 10].

Report Drafting & Recommendations: LLMs can generate review comments, meta-reviews, and even suggest accept/reject decisions, yet risk producing superficial, irrelevant, or hallucinated feedback and cannot fully grasp research novelty or nuanced argumentation [1, 2, 4, 5].

Decision Making: Some tools aggregate multiple reviews and support editors, but there is little documentation of AI making final decisions—human oversight is universal in current workflows [1, 4].

Tool Examples: Statcheck, Penelope.ai, UNSILO, StatReviewer, and custom transformer-based models are repeatedly cited as deployed or prototyped [3, 4, 9, 10].

2. Human vs. AI: Advantages and Limitations

Advantages: Efficiency: Significant reduction in editor/reviewer workload and faster processing times [2, 6, 10]. Detection: Effective at picking up formatting errors, plagiarism, and some aspects of guideline adherence [1, 5, 7, 9]. Reviewer Discovery: AI finds lesser-known or out-of-network reviewers [10].

Limitations: Contextual Judgment: AI struggles with evaluation of novelty, nuanced critiques, and complex argumentation [1, 3, 5]. Bias and Fairness: AI can propagate or amplify existing biases and lacks consistent transparency [1, 3, 7, 10]. Accountability and Hallucinations: Fabrication of references and superficial analysis demand strict human oversight [1, 2, 4].

3. Ethical, Legal, and Institutional Guidelines

Risks Identified: Confidentiality/privacy breaches, especially when using commercial LLMs for unpublished manuscripts [1, 2, 4, 6, 10]. Lack of clarity on responsibility and disclosure obligations for AI use by reviewers [1, 2]. Potential for automation bias, deskilling, and reviewer recognition issues [3, 7, 10].

Guidelines and Policies: Major organizations (COPE, ICMJE, NIH, ARC, leading journals) have issued or updated policies—almost entirely focused on author use, with few on reviewer-side AI, leaving a regulatory vacuum [1, 2]. EU AI Act frames editorial AI systems as «high-risk», but few papers address compliance in depth [3]. Studies call for enforceable, transparent editorial policies including sanctions, disclosure mandates, and human-in-the-loop procedures [1, 2, 10].

4. Author and Reviewer Perspectives

Author Behavior: No direct empirical evidence on changes in submission patterns or trust due to AI—but surveys and perspectives speculate faster processes might attract submissions, tempered by concerns about transparency and fairness [3, 6, 7, 8].

Reviewer Survey Data: Practitioners recognize AI's utility but remain cautious about over-reliance and data privacy; recommend human–AI balance [8, 10].

Recognition and Motivation: Risk of undervaluing human expertise and inadequate credit for mixed AI-human workflows [7, 10].

5. Regional/International Comparison

Papers acknowledge regional regulatory differences and cite the EU AI Act's high-risk classification for peer review AI; few studies assess real-world compliance, and no cross-national empirical comparisons are found [2, 3].

GAPS AND LIMITATIONS IN CURRENT EVIDENCE

Empirical Gaps: Real-world studies covering all peer review stages are scarce—most evidence addresses screening and reviewer assignment; end-to-end deployment and impact on author behavior are almost absent [1, 2, 4]. Virtually no published systematic evaluation of AI's impact on final editorial decisions, long-term trust, or submission volumes.

Conceptual/Policy Gaps: Lack of robust, enforceable institutional or global guidelines for reviewer-side AI use; policy largely focuses on author tools [1, 2, 3]. Ethics and compliance with «high-risk» frameworks (e.g., EU AI Act) not substantively addressed in practice [3, 4].

RECOMMENDED INSTITUTIONAL ACTIONS (SYNTHESIZED FROM REFERENCES)

Develop explicit editorial policies governing reviewer-side and editorial-AI use, including transparency, disclosure, consent, privacy, and sanction mechanisms [1, 2, 10].

Maintain essential human oversight and decision-making at all stages; avoid full automation, especially for nuanced judgment [1, 3, 5].
Implement ongoing reviewer/editor training and recognition systems tailored for AI-augmented workflows [3, 7].
Foster international dialogue on harmonizing guidelines and compliance for «high-risk» AI applications in peer review [2, 3].

CATEGORIES OF RESOURCES FOR «INFLUENCE OF AI IN ALL STAGES OF ACADEMIC PEER REVIEW»

1. Comprehensive Overviews Addressing Multi-Stage AI Integration in Peer Review

Papers that explicitly survey AI's role across the screening, reviewer assignment, reporting, and decision-making stages alongside ethical/legal discussion.
References: [1, 4, 5, 7, 3].
Details: [1]: Reviews AI use in screening, reviewer selection, review/report drafting, and decision letters. Discusses strengths/weaknesses, privacy, bias, lack of policy. [4]: Surveys NLP/AI at each review stage (generation, aggregation, meta-reviews), flags lack of deployment data, discusses bias/human-in-the-loop. [5]: Evaluates AI tools for screening, plagiarism, format, draft assessment; discusses limits in assessing novelty, ethics, privacy; recommends practical guidelines. [7]: Dual focus on drafting and peer review; discusses screening acceleration, bias risks, reviewer recognition, integrity metrics, ethics. [3]: Conceptual overview covers workflow, tools for screening/assignment, comparative strengths/weaknesses, future scenarios, ethics/high-risk regulation.

2. Empirical Evaluations of AI Tools/Platforms in Peer Review Workflows

Papers presenting prototype systems, tested tools, or real-world mixed-method evaluations of AI's effects (e.g., reviewer assignment, screening performance).
References: [9, 10, 2].
Details: [9]: Describes/benchmarks screening and assignment AI tools (Statcheck, Penelope.ai, custom ML), tests on ~3,300 submissions, discusses detected bias, model metrics. [10]: Mixed-methods field study of AI-assisted reviewer selection by 20 editors across disciplines; analyzes efficiency, overlap, bias; qualitative/quantitative results. [2]: Summarizes deployments of LLMs in review authorship, sentiment/tone analysis on large reviewer corpora, effects on report style/turnaround.

3. Ethical, Legal, and Regulatory Analysis / Policy Summaries

Papers with in-depth coverage or summary of ethical challenges, privacy, bias, responsibility, and current institutional/international guidelines (COPE, ICMJE, EU AI Act, etc.)
References: [1, 2, 3, 5, 7, 8, 4, 10, 6].
Details: [1]: Extensive discussion of missing policies, risks in confidentiality, bias, proposals for enforceable oversight. [2]: Catalogs major policy/guideline shifts (NIH, JAMA, Science), highlights journal restrictions, outlines top ethical issues. [3]: Explores EU high-risk AI regulation, future scenarios, reviewer deskilling, fairness impacts. [5]: Reviews transparency, privacy, accountability, overuse risks; sets recommendations for frameworks. [7]: Automation bias, reviewer deskilling, fairness, need for recognition/incentives. [8]: Discusses survey opinions on privacy, trust, need for balance/oversight. [4]: Notes lack of compliance assessment, flags

need for regulations and human oversight. [10]: Addresses algorithmic bias, privacy, and need for ethical frameworks emerging from empirical study. [6]: Advocates journal-restricted corpora for privacy, warns of LLM data issues, but with shallow coverage.

4. Comparison of Automated vs. Human Peer Review: Strengths, Limits, Objectivity

Resources that directly analyze, compare, or summarize differences in objectivity, speed, thoroughness, error/bias detection.
References: [1, 4, 5, 6, 7, 3, 2, 10, 9].
Details: [1]: Faster, more consistent, but superficial, potentially biased and error-prone feedback from AI. [4]: LLMs efficient in generating/aggregating reviews but lack contextual/novelty judgment. [5]: AI checks speed/format; lacks scientific depth, over-reliance risk. [6]: 57–82 % AI review feedback rated «helpful»; overlap with humans substantial at triage but lacks full depth. [7]: Automation increases speed, but risks bias, reviewer deskilling, undervaluation. [3]: Outlines comparative table—AI's speed/objectivity vs. loss of subtlety/expertise. [2]: Notes tone improvement, politeness, speed post-LLM, but reliability remains concern. [10]: Reviewer assignment—AI helps find novel matches; editors flag context/disciplinary gaps; requires human check. [9]: Screening/assignment AI catches some errors, flags bias; doesn't capture all essentials.

5. AI Tools/Platforms Used in Peer Review

Papers identifying/summarizing major tools and computational approaches (Statcheck, Penelope.ai, LLMs, reviewer assignment algorithms, etc.).
References: [9, 3, 4, 5, 7, 2, 10].
Details: [9]: Lists Statcheck, Penelope.ai, UNSILO, StatReviewer, plus custom ML for screening and matching. [3]: Enumerates Kousha & Thelwall, Checco et al.'s tools, AcademicGPT, etc., for filtering/plagiarism/assignment. [4]: Reviews transformer/NLP/LLM approaches (OpenReview-based), meta-review prompting tools. [5]: General mention of grammar/language/plagiarism tools; no specific toolbench. [7]: Focus on transformer-based models for screening/triage; ties to authoring support as well. [2]: Cites GPT, ChatGPT, generative models applied to review/comment creation and analysis. [10]: AI-driven reviewer recommendation system, overlap/efficiency vs. manual process.

6. Perspectives, Theoretical Analyses, and Expert Opinions

Conceptual and position articles offering reflections, scenarios, or recommendations but limited/absent empirical data.
References: [3, 5, 6, 7, 8, 11, 4].
Details: [3]: Presents strategic roadmapping, future scenarios, guidelines adaptation. [5]: Surveys scientific community perspectives, focuses on support-not-replacement stance. [6]: Opinion on future direction, need for restricted corpora, limited metrics. [7]: Advocates incentives, authenticity measures; largely descriptive. [8]: Survey-based opinions/sentiments from 685 practitioners. [11]: Pure editorial/opinion on potential/risks of LLM-assisted peer review. [4]: Concludes with need for greater compliance and oversight, highlights research gaps.

7. International/Regional Guidelines, Institutional Recommendations, and Comparison of Regulatory Approaches

Coverage of COPE, ICMJE, NIH, EU AI Act, differences in practices between major regions/journals.
References: [1, 2, 3, 5, 8].

Details: [1]: Reviews lack of AI policies for reviewers vs. authors by major bodies (COPE/WAME), urges new procedures. [2]: Summarizes NIH, ARC, JAMA, Science, The Lancet, ICMJE positions—banning, restricting, or requiring disclosure of AI in review. [3]: Discusses EU's AI Act «high-risk» category for peer review. [5]: Mentions need for institutionally enforceable frameworks/guidelines but limited comparative depth. [8]: Global survey, remarks on rising digital norms and transparency policies.

8. Surveys and Mixed-Methods Studies of Stakeholder Perceptions, Author/Reviewer Attitudes, and Trust

Empirical resource on peer reviewer or editor attitudes, author perceptions, AI's impact on trust/incentives.
References: [8, 10].
Details: [8]: Global survey/analysis of 685 researchers on trust and AI in peer review; concerns about privacy and the balance of digital/human models. [10]: Editor feedback on AI reviewer recommendations—fit, suitability, time savings, bias, operational performance.

9. Documentation of Journal Experiences/Case Studies (Emblematic or Pioneer Journals)

Papers or reports that highlight practical experiences or pilot deployments by specific journals.
References: [3, 2, 10].
Details: [3]: Details editorial workflow and AI considerations at Revista Española de Sociología. [2]: Summarizes experiences of neuroscience journals with AI-generated review comments; cases from multiple high-impact journals. [10]: Practical trial of AI reviewer-assignment in various journals, including STEM and other fields.

10. Coverage (or not) of Author Submission Behavior and Willingness to Submit Under AI Review Systems

Whether resources empirically address (or explicitly note the lack of data on) author-side behavioral impacts.
References: [4, 5, 6, 7, 8]
Details: [4]: Notes no studies report on author submission behavior or trust under AI-reviewed systems. [5, 6, 7, 8]: Similarly, mention the importance of the issue but do not provide empirical findings—this is a notable research gap.

TIMELINE

Figure 1 shows the publications obtained in the UnderMind-assisted review (the color shows the relevance of the publication with respect to the topic – in gray the maximum relevance, in gray the minimum relevance – measured by the reference rate, which is the fraction of relevant articles that cite a resource).

Timeline and Historical Development

Early Prototypes and Analytical Frameworks (2021–2022): [9] (2021) presented one of the first empirical prototypes for AI-assisted peer review, focusing on screening, reviewer assignment, and ethical concerns (bias, transparency). Cited widely, it established foundational concerns and tool categories (Statcheck, Penelope.ai). Conceptual discussions around screening and reviewer matching began appearing, but empirical multi-stage deployments were scarce.
Theory-Driven Expansion, Ethics, and Preliminary Guidelines (2023): [3, 7] (2023) expanded the conversation with reflection on AI's transformative potential, scenario planning (future pathways for IA/peer review), and ethical issues (deskilling, over-automation, bias). Papers started to discuss

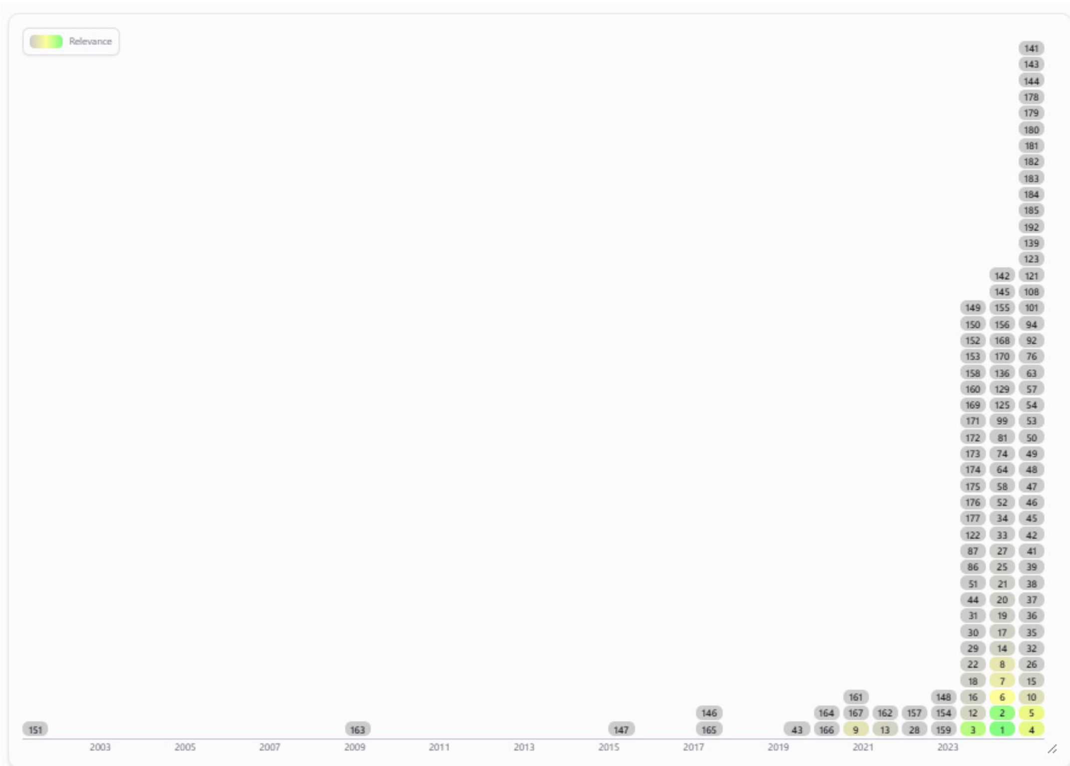

Figure 1. Top References Over Time about n=192 articles reviewed. Bibliographic citations are available in the supplementary Excel document.

the regulatory environment (EU AI Act) and institutional tensions but mostly lacked empirical deployment evidence. [7] Also explicitly linked operational risks (reviewer undervaluation), started framing incentive/recognition questions, and flagged the need for human-in-the-loop designs.

Operationalization, Comparative Studies, and Institutional Policy Focus (2024): [1, 2, 6, 8, 10, 11] (2024) document a surge in mixed-methods and empirical studies: [10] Conducted a real-world, cross-disciplinary evaluation of AI reviewer selection (time savings, overlap/suitability metrics, bias/privacy). [2] Analyzed LLMs (ChatGPT, GPT-4) used for both review drafting and linguistic analysis of large review corpora, highlighting shifts toward positive/polite tone and emerging institutional rules/policies (e.g. COPE, ICMJE, Science, JAMA stances on disclosure/confidentiality). [1] Examined the policy gap for reviewer AI use versus existing policies for authors, advocating enforceable, transparent guidelines and linking AI use across all peer review phases to integrity risks. [6] reviewed case examples and early empirical evidence for LLM-assistance (overlap with human critique), while emphasizing persistent risks (hallucinations, confidentiality, fairness). [8] Reported on community perception

(685 survey respondents) regarding digital/AI tool adoption and integrity, with an emphasis on continued trust, flexibility, and privacy concerns. [11] Provided a conceptual perspective on generative AI, focusing mainly on pitfalls (bias, hallucinations, ethics) but without broad empirical coverage.

Breadth, Domain Expansion, and Survey Synthesis (2025, early online/preprints): [4, 5] (2025) provide comprehensive surveys/overviews: [4] Reviewed multi-stage LLM-powered peer review and meta-review automation, mapping data gaps (domain diversity, author reactions, guideline compliance), and compiling lessons from analyses of large OpenReview corpora. [5] Consolidated ethical/methodological debate, arguing for balanced human–machine integration and issuing practical guidelines for preserving integrity alongside efficiency. Both noted that real-world deployments remain limited outside of conference and STEM contexts, and that empirical evaluation of downstream effects (author behavior, decision consistency, regulatory impact) is still largely missing.

Key Research Clusters and Individual Contributions

Checco et al. and AI-Peer Review Systems [9], cited by [1, 3, 4]: The group led by Alessandro Checco introduced and tested one of the earliest functioning AI peer review workflows, establishing benchmarks for initial screening, automated scoring, and reviewer–manuscript matching. Their work is the empirical anchor for many subsequent conceptual and survey-based analyses.

Policy and Editorial Practice Focus (Bauchner & Rivara [6], Mollaki [1], Farber [10]: Howard Bauchner and F. Rivara [6] are editors and policymakers who catalyzed the practical/policy-facing debate on LLM integration, focusing on how journals should adapt screening/triage workflows. V. Mollaki [1] foregrounds the lack of standardized policy for reviewer-side AI (vs. author AI use), highlighting the need for explicit protocols and accountability mechanisms. Shai Farber [10] contributed the most robust mixed-methods study on AI-aided reviewer selection, providing rare large-scale quantitative evidence and nuanced qualitative editor perspectives/incentives.

Scenario Planning and Ethical Threat Assessment (Cárdenas [3], Plebani et al. [7]: Julián Cárdenas [3] offered a broad scenario framework for the sociotechnical evolution of AI in peer review, linking technical and regulatory shifts, especially in Spanish and EU regulatory contexts. Mario Plebani's group [7] foregrounded ethical implications (deskilling, reviewer undervaluation, incentive structures) and mapped both paper drafting and review assistance from a policy/operational standpoint.

Synthesizers and Surveyors [4, 5]: Eger et al. [4] and Doskaliuk et al. [5] provided recent, thorough summaries and meta-analyses, consolidating foundational work to clarify open questions, methodological boundaries, and roadmaps.

Essential Takeaways

Empirical foundation for AI-assisted peer review was established by Checco et al. [9] in 2021, later expanded into mixed-methods real-world testing by Farber [10] and broad reviews/meta-analyses [4, 5] in 2024–2025.

The main conceptual and ethical advances (scenario planning, regulatory analysis, incentive structure debate) come from Cárdenas [3], Plebani [7], and policy-focused research [1, 6].

Recent years show a shift: from isolated tool-building and high-level reflection (2021–2022) to more grounded, policy- and practice-driven empirical research and interdisciplinary surveys (2023–2025).

No single «dominant lab» but influential clusters include the AI–publishing crossover (Checco, OpenReview/NLP researchers), editorial policy and regulatory experts (Bauchner, Mollaki, Farber), and scenario planners (Cárdenas, Plebani).

Key open gap remains—evidence of AI impact on author submission behavior, final editorial decision process, cross-journal comparisons, and adherence to international regulatory standards.

REFERENCES

1. Mollaki V. Death of a reviewer or death of peer review integrity? The challenges of using AI tools in peer reviewing and the need to go beyond publishing policies. Research Ethics. 2024; 20(2):239-250. Accessed on 26/04/2025. Available at: https://doi.org/10.1177/17470161231224552
2. Cheng K, Li C, et al. Generative artificial intelligence is infiltrating peer review process. Critical Care. 2024; 28:149. Accessed on 26/04/2025. Available at: https://doi.org/10.1186/s13054-024-04933-z
3. Cárdenas J. Artificial intelligence, research and peer review: future scenarios and action strategies. Spanish Journal of Sociology. 2023; 32(4):49-54. Accessed on 26/04/2025. Available in: https://doi.org/10.22325/fes/res.2023.184
4. Eger S, Miller T, et al. Transforming Science with Large Language Models: A Survey on AI-assisted Scientific Discovery, Experimentation, Content Generation, and Evaluation. ArXiv. 2025; 2502.05151. Retrieved 26/04/2025. Available at: https://doi.org/10.48550/arXiv.2502.05151
5. Doskaliuk B, Yatsyshyn R, et al. Artificial Intelligence in Peer Review: Enhancing Efficiency While Preserving Integrity. Journal of Korean Medical Science. 2025; 40(7):e92. Accessed on 26/04/2025. Available at: https://doi.org/10.3346/jkms.2025.40.e92
6. Bauchner H, Rivara F. Use of artificial intelligence and the future of peer review. Health Aff Scholar. 2024; 2(5):QXAE058. Accessed on 26/04/2025. Available at: https://doi.org/10.1093/haschl/qxae058
7. Carobene A, Plebani M, et al. Rising adoption of artificial intelligence in scientific publishing: evaluating the role, risks, and ethical implications in paper drafting and review process. Clinical Chemistry and Laboratory Medicine (CCLM). 2023; 62(5):835-843. Accessed on 26/04/2025. Available at: https://doi.org/10.1515/cclm-2023-1136
8. Calamur H, Ghosh R. Adapting peer review for the future: Digital disruptions and trust in peer review. Learned Publishing. 2024; 37:49-54. Retrieved 26/04/2025. Available at: https://onlinelibrary.wiley.com/doi/epdf/10.1002/leap.1594
9. Checco A, Bianchi G, et al. AI-assisted peer review. Humanities and Social Sciences Communications. 2021; 8(1):25. Accessed on 26/04/2025. Available at: https://www.nature.com/articles/s41599-020-00703-8
10. Farber S. Enhancing peer review efficiency: A mixed-methods analysis of artificial intelligence-assisted reviewer selection across academic disciplines. Learn Publ. 2024. Accessed on 26/04/2025. Available at: https://doi.org/10.1002/leap.1638