J Korean Med Sci. 2025 Dec 22;40(49):e342
https://doi.org/10.3346/jkms.2025.40.e342
eISSN 1598-6357·pISSN 1011-8934

JKMS

Review Article
Editing, Writing & Publishing

Artificial Intelligence in Detecting Statistical Errors: Implications for Authors, Reviewers, and Editors

Fatima Alnaimat,[1] **Abdel Rahman Feras AlSamhori**,[2] **Husam El Sharu**,[3] **Leen Othman**,[2] **Aizhan Oralbek**,[4] **and Olena Zimba** [5,6,7]

[1]Division of Rheumatology, Department of Internal Medicine, School of Medicine, University of Jordan, Amman, Jordan
[2]School of Medicine, University of Jordan, Amman, Jordan
[3]Department of Rheumatology at Case Western University/University Hospitals, Cleveland, OH, USA
[4]Department of Social Health Insurance and Public Health, South Kazakhstan Medical Academy, Shymkent, Kazakhstan
[5]Department of Rheumatology, Immunology and Internal Medicine, University Hospital in Kraków, Kraków, Poland
[6]National Institute of Geriatrics, Rheumatology and Rehabilitation, Warsaw, Poland
[7]Department of Internal Medicine N2, Danylo Halytsky Lviv National Medical University, Lviv, Ukraine

Received: Oct 26, 2025
Accepted: Nov 25, 2025
Published online: Dec 9, 2025

Address for Correspondence:
Fatima Alnaimat, MD
Department of Internal Medicine, Division of Rheumatology, School of Medicine, University of Jordan, Amman 11942, Jordan.
Email: f.naimat@ju.edu.jo

ORCID iDs
Fatima Alnaimat
https://orcid.org/0000-0002-5574-2939
Abdel Rahman Feras AlSamhori
https://orcid.org/0000-0002-2715-4320
Husam El Sharu
https://orcid.org/0000-0003-4354-7236
Leen Othman
https://orcid.org/0009-0006-1462-5601
Aizhan Oralbek
https://orcid.org/0009-0003-2616-4660
Olena Zimba
https://orcid.org/0000-0002-4188-8486

ABSTRACT

Choosing the right statistical tests is essential for reliable results, but errors, like picking the wrong test or misinterpreting data, can easily lead to incorrect conclusions. Research integrity implies presenting research that is honest, clear, and uses correct statistics. By identifying statistical errors, artificial intelligence (AI) systems such as Statcheck and GRIM-Test increase the reliability of research and assist reviewers. AI helps non-experts analyze data, but it can be unpredictable for experts dealing with complex data analysis. Still, its ease of use and growing abilities show promise. Recent studies show that AI is increasingly helpful in research, assisting in spotting errors in methodology, citations, and statistical analyses. Tools like LLMs, Black Spatula, YesNoError, and GRIM-Test improve accuracy, but they need good data and human checks. AI has moderate accuracy overall but performs better in controlled settings. The Statcheck and GRIM-Test are especially good at spotting statistical errors. As more studies are retracted, AI offers helpful, albeit imperfect, support. It can speed up peer review and reduce reviewer workload, but it still has limits, such as bias and a lack of expert judgment. AI also brings risks like misreading results, ethical issues, and privacy concerns, so editors must make final decisions. To use AI safely and effectively, large, well-labeled datasets, teamwork across fields, and secure systems are required. Human oversight is always necessary to review research processes and ensure their reliability; humans must make the final decision and utilize AI responsibly.

Keywords: Artificial Intelligence; Publications; Statistics; Scientific Misconduct

INTRODUCTION

Research integrity (RI) refers to adherence to the ethical standards, emphasizing intellectual honesty and transparency across all stages of the research process. The opposite end entails fabrication, falsification, and plagiarism. RI allows drawing meaningful results from

Disclosure
The authors have no potential conflicts of interest to disclose.

Disclosure of AI Use
Grammarly (https://app.grammarly.com/) was utilized to enhance the grammar and clarity of this manuscript. All authors have reviewed AI edits and revised them multiple times.

Author Contributions
Conceptualization: Alnaimat F, Oralbek A, Zimba O. Investigation: Alnaimat F. Supervision: Alnaimat F. Visualization: Alsamhori ARF. Writing - original draft: Alsamhori ARF, El Sharu H, Othman L. Writing - review & editing: Alnaimat F, Alsamhori ARF, El Sharu H, Othman L, Oralbek A, Zimba O.

reliable findings, maintains confidence in research, and prevents adverse impacts on the public.[1] The inconsistent RI has been pointed out by the failure of reproducibility in scientific research, which has become increasingly recognized, resulting in a surge in the number of research retractions.[2] Among the reasons for research retractions is a decline in statistical integrity. Statistical integrity implies obedience with the statistical code of conduct, which includes objectivity, accuracy of primary data, and the use of appropriate statistical techniques, thereby preventing the drawing of incorrect conclusions.[3]

The application of sound statistical tests is a vital part of original studies and meta-analyses. Statistical errors can lead to false-positive or false-negative interpretations during data analysis and influence the validity, credibility, and reproducibility of research findings. Common statistical errors in published research include the inappropriate use of the *t*-test and Pearson's bivariate correlation, errors in logistic and linear regression analyses, the incorrect reporting of *P* values solely, and the incorrect reporting of medians and means and interquartile ranges. In meta-analyses, additional errors surface, such as inappropriate model selection, biased study selection, dependency, and ignoring incomplete data in selected studies. These errors place a burden on reviewers and journal editors when trying to ensure the methodological soundness of the paper at hand.[4] In a literature review of psychology articles from 1985 to 2013, Nuijten et al.[5] reported an inconsistency rate of nearly 10.6% in the reporting of null hypotheses and *P* values. Unfortunately, this accelerated in the coronavirus disease 2019 (COVID-19) era. Jung et al.[6] measured the adherence to methodological tools using the Cochrane risk bias tool. They detected a 32% decline in the adherence rate to a high methodological quality.

Artificial intelligence (AI) has been increasingly utilized in academic research for tasks including idea generation, content improvement, literature review support, and data analysis enhancement.[7] This growing role presents a significant paradox.[7] On one hand, tools like ChatGPT has been presented as a cost-free solution that could dismantle the "linguistic barrier," helping non-Anglophone authors from regions with moderate to low economic stability achieve parity with native-speaking counterparts.[8] There is now an increasing emphasis on the role of AI in helping with statistical review. AI has been increasingly utilized in academic research, including idea generation, content improvement, literature review support, and data analysis and management enhancement. Unfortunately, such use has resulted in a breach of RI. In a systematic review by Khalifa and Albadawy,[9] 24 studies were identified examining the impact of AI on academic research. While many of these studies highlighted AI's usefulness, particularly in scientific writing, several raised concerns regarding RI and the potential for inaccuracies. This past decade has also witnessed the development of computational tools like GRIM-Test and Statcheck, which can scan submitted papers and detect internal inconsistencies with *P* values and means. In addition, language models like ChatGPT can serve as an objective critique of the methods and findings in papers.[10,11] However, AI is essentially a machine that responds to a set of data inputs fed into it, which limits its role in accessing raw data not included in the submitted paper. In addition, complex statistical methods like Bayesian Analysis, mixed effect models, or survival analysis usually exceed the capacity of AI tools.

The increasing use of AI in article submission necessitated further use by reviewers and editors. Several platforms have used AI to assist in the initial reviewing process and in facilitating peer review.[12] Although this use of AI has already been established, using AI in proofreading methodological and statistical integrity has not been well reviewed.

This review aims to explore the possibility of using AI tools in detecting statistical mistakes and errors in published research articles, with particular focus on how journal editors and reviewers can utilize such tools so that they become assistants in the review process. Ultimately, we aim to support the development of standardized criteria that can help in guiding decisions about article retraction or correction.

TYPES OF STATISTICAL MISTAKES AND ERRORS IN PUBLISHED LITERATURE

Statistical tests must be applied accurately, depending on the type of data, to yield desirable outcomes. Statistical analyses can be misapplied in several ways, leading to inaccurate results and misleading conclusions.[3] Common examples include[13,14]:

- Errors in statistical design: inappropriate assumptions while selecting a statistical test.
- Errors in data presentation: inappropriate communication of data, either with visual designs, labels, or descriptions.
- Errors in data outlier treatment: inappropriate exclusion of outliers with the assumption of having minimal impact on results.
- Errors in the use of parametric and nonparametric tests: inappropriate use of a parametric test with nonparametric data or vice versa.
- Errors in the disclosure of post-hoc analyses: the analysis of data after data collection in a post-hoc form without disclosing such analysis.
- Errors in defining the alternative and null hypotheses.

The increasing incidence of data misuse is likely related to the lack of statistical knowledge. In a cross-sectional study conducted by Gore et al.[15] among faculty members and students from various medical colleges, more than 50% of participants reported finding statistics difficult and struggled to describe the meaning of *P* values, standard deviation, and appropriate sample size. On the other hand, many published articles fail to discuss statistical assumptions, specify the exact statistics used in analysis, and employ vague terms when reporting statistical tests.[14]

In addition to misleading reporting, errors may also arise during data interpretation. Examples include presenting only the odds ratio instead of both the odds ratio and risk difference for dichotomous outcomes, as the latter improves interpretability, and failing to describe how missing data were handled. Moreover, interpreting I^2 as an absolute measure of heterogeneity, similar to T^2 in meta-analysis, can lead to misinterpretation.[16]

An example of the serious effects of data interpretation was the 'pill scare' in the United Kingdom (UK). The reporting of a two-fold increase in thrombosis events in women who are taking oral contraceptives, rather than reporting an increase from about 1 in 7,000 to 2 in 7,000, resulted eventually in an increase in the pregnancy and abortion rates in the UK in 1995.[17]

Reproducibility, the ability to regenerate similar data with the same methods of analysis, has various types. These include reproducing experimental conclusions with either the same or a different analysis method, or reproducing the same conclusion with a new dataset if the same laboratory was used.[18] With the declining rate of RI, the ability to reproduce medical research has gained attention in the medical community. This usually stems from

https://doi.org/10.3346/jkms.2025.40.e342

inappropriate study design or baseline comparisons, and miscalculation. Sometimes, reproducibility is not feasible given a vague description of the methodology.[19]

OVERVIEW OF AI FOR STATISTICAL ANALYSIS

With AI evolution, it has also been widely used in performing statistical analysis, especially among less-skilled users, by providing step-by-step solutions for complex statistical tasks.[20] Although some AI models have been fine-tuned for statistical tasks, the building of AI is based on training using mainly linguistic and language datasets, rather than statistical codes.[21] When challenged with complex analyses, AI performance varies. Wetzelhütter and Prandner[20] tested the reproducibility of published analyses by asking AI to generate the syntax, running the analysis, and exploring the results. Unfortunately, AI was only able to provide basic syntax and fragments that were tangentially related to the problem, which represents an issue for naïve users.[22] Methodologically, this study by Wetzelhütter and Prandner,[20] they effectively highlight AI's current limitations, their soundness is not that of a large-scale validation trial.[20] They demonstrate feasibility and inconsistency rather than providing robust performance benchmarks.[20] When comparing AI to the current statistical programs, AI produces similar results in simple analyses such as the sample *t*-test, paired *t*-test, simple linear regression, and Pearson and Spearman's correlations. However, discrepancies in post-hoc analyses, confidence intervals, and more complex tests were noted.[23]

Given the AI nature of training, it showed equivalent consistency of results in descriptive results compared to commonly used statistical packages (SAS, SPSS, and R). Notably, when asking AI to follow the R software, it produced better outcomes, likely because R is an open-source software. On the other hand, comparing AI with those packages in Intergroup correlational and correlational analyses showed statistical discrepancies.[24]

Despite all inconsistencies, AI has an exceptional user-friendliness compared to other packages with growing capabilities that are comparable to such programs and potentially competing with them in the near future.

AI PROGRAMS USED FOR DETECTING STATISTICAL MISTAKES AND ERRORS

Considering the growing popularity of AI tools and their increasing role in research, many articles have been recently published in discussion of this topic. A 2025 article published in Nature discussed the use of AI in reviewing errors, methodology, calculations and inconsistencies in the citation of papers.[25] Another 2025 article in the Frontiers explored the use and possible risks that accompany the use of AI tools in the drafting of paper manuscripts and in data analysis. It also explored the use of tools like large language models (LLMs) in the detection of reference errors, the Black Spatula Project and YesNoError tools in the detection of mathematical and experimental logic errors.[26] Another article shed focus on explaining the use of the GRIM-Test as an AI tool for the detection of statistical errors, it demonstrates the simplicity of the tool allowing for its use not just by skilled reviewers and editors, but by authors too, and it highlights the importance of using accurate raw data when employing the tool because it is a major factor in determining the outcome of its use.[10]

https://doi.org/10.3346/jkms.2025.40.e342

EVALUATION AND VALIDATION OF AI TOOLS

Several studies evaluated the AI diagnostic accuracy. For instance, a meta-analysis by Takita et al.[27] evaluated studies between 2018–2024, and concluded that among different AI models, AI had a diagnostic accuracy of 52% (95% confidence interval, 47.0–57.1%). However, AI error detection has not been well studied in medical practice. In a study that applied video-assisted operating rooms (OR), AI helped in identifying vial swap errors with a sensitivity and specificity of 99.6% and 98.8%, respectively.[28] In a less controlled setting compared to the OR, the AI error-reduction rate ranged from 40–55%.[29] When applying AI error detection to medical research, various AI platforms are available to identify research-related errors, whether in bias detection, plagiarism, or proofreading.[12] However, the exact sensitivity and specificity of AI research detection abilities are not documented in the literature.

Recent retractions of published papers have highlighted the types of errors that can be detected using AI tools if they were to be used in these studies. A 2004 systematic review declared poorer clinical outcomes of chronically ill patients using interactive health communication applications. It was later discovered that the study contained critical errors in data analysis that invalidated the findings.[30] It is retractions like this one that highlight the increasing value of using AI tools as assistants in the detection of statistical and reporting inconsistencies.

However, these same tools are not devoid of limitations in terms of their accuracy, sensitivity, and specificity. For example, Statcheck is limited to scanning APA-style statistics, and it would not be able to recognize data organized in tables. Still, according to a study that put its performance in comparison to manual coding, the sensitivity was found to be 85.3–100%, the specificity was found to be close to 96–100%, and its overall accuracy is 96–99.9% indicating the reliability of this tool.[11] The GRIM-Test is a mathematical tool that ensures reported means are possible given a certain sample size. It has been proposed as a reliable tool by validity studies in which its sensitivity exceeded 83%, specificity was over 96%, and overall accuracy reached above 92%.[31] Methodologically, the validation of GRIM is robust because it is based on a fixed mathematical principle rather than interpretive analysis.[10] The soundness of its validation studies therefore depends on the diversity and size of the corpus of papers used for testing.[10] A methodological weakness, as noted, is that the tool's accuracy is entirely dependent on the clear and accurate reporting of statistics in the source paper.[10] In addition, GRIM and GRIMMER (an extension of GRIM) can both be implemented in R or Python to detect statistical errors. This has allowed for the automation of this process, making it more time efficient. Nonetheless, the accuracy and sensitivity of this method are limited to the clear outline of the statistics in a paper, allowing the algorithm to detect the errors correctly. The specificity, at the same time, depends greatly on the way that the tool was built, because the use of either strict or oversimplified rules can cause the incorrect flagging of correct statistics as errors.[32] Despite this available information, there is still a paucity in studies that sufficiently explore the extent to which these tools could be used to date.

On the other hand, research retraction has reached a 'crisis' level in the context of the pressure of publishing. Taylor and Francis had to retract 350 papers in 2022 alone, while Wiley retracted more than 8,000 articles in 2023.[33] Among the reasons for those retractions are data and analysis-related retractions. In a study that collected almost 50,000 retractions in 2023, nearly 19–32% of those retracted articles were related to data handling and analysis.[34,35] In the era of COVID, *The Lancet* journal has retracted a study about the use of

https://doi.org/10.3346/jkms.2025.40.e342

hydroxychloroquine in COVID, using the 'Surgisphere' data source, and the following analysis could not be validated.[36]

In the context of the increasing number of research retractions, few AI platforms have been developed to screen articles for potential retraction, which can also be used in the peer-review process to evaluate the possibility of future retraction.[37] **Table 1** provides a concise overview of the mentioned AI tools.

A synthesis of the tools presented in **Table 1** reveals distinct categories and a significant trade-off between specificity and scope. The tools fall into two main groups: highly specific algorithms (like Statcheck, GRIM, and GRIMMER) and generalist critique models (like LLMs).[10,11]

The specific algorithms offer the primary advantage of high, verifiable accuracy for a narrow, objective task.[39] For example, the Statcheck excels at one specific thing: verifying the consistency between *P* values and their associated test statistics in APA format.[11,38] GRIM is similarly narrow, checking the mathematical plausibility of means given a sample size.[10,31,32] Their strength is their reliability; their weakness is their limited scope. They cannot, for instance, assess the appropriateness of a chosen statistical test or interpret data presented in tables.

LLMs represent the opposite end of this trade-off.[10,11,21,23] They can provide a broad, qualitative critique of an entire methodology section or statistical plan, something the specific tools cannot.[10,11,21,23] However, their methodological soundness is far less proven.[10,11,21,23] As noted, their performance is inconsistent on complex analyses, they risk "hallucinating" or misinterpreting information, and their "black-box" nature makes their reasoning difficult to verify.[40]

Therefore, this review's synthesis suggests that no single AI tool is a complete solution for statistical review. A robust AI-assisted workflow would not rely on one tool but would use them in combination: employing specific, high-accuracy tools like Statcheck and GRIM to screen for numerical and reporting-level errors, while cautiously using LLMs to flag potential qualitative issues in the study design, all of which must be confirmed by an expert human reviewer.[10,11,41]

Table 1. Summarizing data

AI tool	Primary function	Reported accuracy (sensitivity/specificity)	Key limitations	References
Statcheck	Automatically scans reported statistics (APA-style) to detect inconsistencies in *P* values and test results	Sensitivity 85–100%; Specificity 96–100%; Overall accuracy ≈ 97–99%	Limited to APA-style text; cannot interpret data in tables; may misread complex tests	11,39
GRIM-Test	Checks internal consistency of reported means against sample size to verify mathematical plausibility	Sensitivity > 83%; specificity > 96%; overall accuracy > 92%	Requires clearly reported means and sample sizes; limited for non-integer or derived data	27,32,33
GRIMMER	Extension of GRIM allowing detection of multiple statistical inconsistencies simultaneously	Comparable accuracy to GRIM; faster automated performance	Relies on explicit numerical reporting; oversimplified thresholds may cause false positives	33
YesNoError	Detects mathematical and experimental logic inconsistencies using algorithmic logic checks	Accuracy not standardized; performs best in controlled data environments	Limited published validation data; prone to over flagging	26
Black Spatula	Identify citation and logical inconsistencies in manuscripts	Qualitative accuracy high in pilot studies; no quantitative benchmark yet	Experimental tool; dependent on text structure and training corpus	26
Large Language Models (e.g., ChatGPT)	Provides general critique of methods, data reporting, and statistical coherence	Comparable to human reviewers in simple analyses; inconsistent in complex models	Lacks access to raw data; risk of misinterpretation and bias	11,21,23,27

https://doi.org/10.3346/jkms.2025.40.e342

HOW AI CAN BE EMPLOYED FOR DETECTING ERRORS AND RETRACTING ERRONEOUS RESEARCH REPORTS AND META-ANALYSES

Technology companies are now signing agreements with publishers to grant them access to articles in machine learning. However, among those articles are retracted articles, which can lead to the spread of misinformation that can be disseminated quickly.[33] On the other side, using retracted articles to train AI to help predict future or possible retractions has also been studied. In a study by Fletcher and Stevenson,[37] who studied the precision of the AI platform Llama 3.2 Base Model, achieving an accuracy of 68%. Other platforms have also been used to detect citation accuracy, plagiarism, bias detection, and image integrity. Other AI-assisted tools have been used to aid in the adherence to journal-specific guidelines. These error-detecting tools can be applied in the submission process and to screen for potential flaws in post-published research.[12]

In a world where AI is becoming an increasingly reliable assistant, journal editors and publishers should no longer have to review papers to detect errors manually. For example, the Statcheck scans the submitted papers and extracts the reported data, recalculates the *P* values, and compares the recalculated values to the values reported in the paper; then, based on that, it flags inconsistencies in the findings.[5] GRIM uses the reported sample sizes and matches them to the total sum divided by the sample sizes in order to detect the plausibility of the reported mean values in a paper. The further automation of these tools is currently done by implementing them in python or R codes.[10]

By automating the detection of statistical errors, AI can significantly reduce the burden on reviewers. By detecting evident errors in papers, the role of reviewers could become more specialized in delving into the complex statistical and methodological faults that are otherwise non-detectable by AI tools; allowing for the quicker detection of papers that should be retracted or corrected by authors. On the long run, AI tools will prove to be invaluable in enhancing the integrity of published research.[38]

AI can be a very time-effective tool that not only saves the time of reviewers and journal editors but can also accelerate the review process, ensuring the completion of a larger number of reviews over a shorter period. In a time when the review process of submitted papers can take months, the use of AI tools as assistants will accelerate the process, possibly to weeks after submission. These tools can scan and process large amounts of text with objectivity, minimizing the window for human error when reviewing long papers.[38]

On the other hand, AI still lacks the in-depth, expert-level subject understanding that is required to critique articles. Additionally, AI could be trained on data sets that have existing biases, where AI may replicate those biases, all of which may influence the accuracy of AI in such tasks.[38] Accordingly, the *Journal of the American Medical Association* (*JAMA*) has introduced specific instructions on using AI in peer review and recommended using it as an assistant.[42]

The research community has resisted AI for various reasons. AI can impact preserving human expertise; it has been suggested that the over-reliance on AI tools negatively impacts critical thinking, decision-making capabilities, and analytical reasoning.[43] When an article gets retracted, it remains flagged as a retracted article and cannot be deleted from the literature,

putting the author's career on the line.[44] Like the use of AI in the peer-review process, AI has been used in reviewing already published articles. However, AI tools function differently and are not equal in their capabilities; thus, making it difficult to predict how AI would screen publications.[45] Given the sensitivity of the retraction process and the potential for false positives when using AI to screen articles for retraction, the medical community became hesitant to use AI for this purpose.[46]

Seeing the limitless possibilities that these tools hold, however, it remains crucial to remember that their use should yet be limited to assist and not fully replace human power. A reasonable, but rather avertable concern regarding AI is the growing dependence of these tools in the scientific field. The limitations to their use should be decided by us, final decisions regarding retraction should remain in the hands of human expertise with the necessary experience. Human touch remains essential in the review of aspects otherwise undetectable by AI tools.

ETHICAL CONSIDERATIONS

When using AI tools, there's going to be room for misinterpretation due to the lack of flexibility that these models possess. Statcheck is known to misinterpret complex statistical tests as errors because it relies on a predetermined model that is incapable of processing tests of such complexity. When errors are detected, a concise and clear explanation cannot always be expected, because AI tools make decisions that might be difficult to interpret. If, and when the decision is left on this sole interpretation, authors will be subjected to unjustified scrutiny that can potentially harm their reputations without a clear and concise explanation that could've otherwise been provided by the editor or reviewer.[47]

Ethical considerations are one of the major limitations in AI use. In the submission period, AI does not fulfill the authorship criteria in holding responsibilities and integrity assurance. Consequently, authors should disclose which part of their research was AI-assisted and hold the responsibility and the originality for their work.[48] Disclosure is deemed unnecessary when AI is employed for tasks like grammatical correction.[7] It is considered unethical to let AI produce significant parts of a paper without revealing this information.[7] In addition, the use of AI to assist in peer review can impose a breach in confidentiality, especially when uploading unpublished materials into AI systems, which may expose sensitive information, violating the peer-review agreements.[49]

When applying AI in the post-publication period, there is a potential for false positives in identifying flaws in published research. Unfortunately, AI does not provide explanations for its decisions, often termed a 'Black-Box Model'; thus, adopting such models in sensitive decisions such as research retraction is difficult.[50] Also, the rapid evolution in AI algorithms makes it more challenging to depend on it at the current time. All of these have raised substantial ethical considerations on using AI in the post-publication checks.

In the article submission process, when editors use AI to evaluate and investigate submitted articles, it is their responsibility to assess the authorship and validity of the article. Similarly, it is the responsibility of the editors to assess the AI-reported claim about a certain article before it reaches the author or the public.[51]

https://doi.org/10.3346/jkms.2025.40.e342

Accordingly, when ensuring that AI tools are only considered assistants in the process of review, the accountability for decisions regarding retraction, corrections, or expressions of concern remain within the hand of human expertise.[52]

FUTURE DIRECTIONS AND RECOMMENDATIONS

AI holds enormous potential. AI is our apprentice, absorbing the skills and patterns we feed into it, and so the accurate detection of errors can be improved by training the tools on large, accurately labeled sets of data. Given the current trends in the sensitivity of current AI models in detecting statistical errors in already published research, better training data must be generated to help facilitate quicker and more precise AI evaluation of statistical errors in submitted and published articles. To do that, a multidisciplinary team of computer scientists and engineers, statisticians, reviewers, and editors is necessary to feed into every aspect of the process.

AI tools might not be as reliable in the meantime, but it won't be long before they become a very reliable assistant. Once a statistical AI model is developed, just like other models, the other concerns would remain unaddressed, including confidentiality breaches, the sensitivity and specificity of the model in different types of statistical errors, and providing explanations for AI-related decisions.

Before manuscript submission, tools can also be utilized by authors to ensure the submission of high-quality papers. However, in the context of clinical research, where sensitive patient information must remain anonymous, how safe are AI tools with such data? How can we ensure that information fed into such machines will never fall into public hands? Before such tools are allowed to review sensitive data, secure systems have to be ensured in order to maintain trust, ethical standards, and data privacy.[52]

CONCLUSIONS

AI can significantly enhance the quality of research by identifying statistical errors and assisting in the identification of papers that may require retraction or correction. Tools are available that can quickly find inconsistencies in data, making research more accurate and reliable while saving time for reviewers and editors. Researchers, editors, and publishers are encouraged to start using these tools responsibly and create clear standards for their use. Training and collaboration are needed to ensure AI is applied correctly and ethically. Still, AI can detect mistakes, it cannot fully understand context or make final ethical decisions. For an optimal balance, AI can serve as a supportive tool while humans perform the final evaluation. This collaboration enhances RI and strengthens trust in scientific publishing.

REFERENCES

1. Zhaksylyk A, Zimba O, Yessirkepov M, Kocyigit BF. Research integrity: where we are and where we are heading. *J Korean Med Sci* 2023;38(47):e405. PUBMED | CROSSREF
2. Miller G, Spiegel E. Guidelines for research data integrity (GRDI). *Sci Data* 2025;12(1):95. PUBMED | CROSSREF

3. Gelfond JAL, Heitman E, Pollock BH, Klugman CM. Principles for the ethical analysis of clinical and translational research. *Stat Med* 2011;30(23):2785-92. PUBMED | CROSSREF
4. Darling HS. Statistical errors in scientific research: a narrative review. *Cancer Res Stat Treat.* 2024;7(2):241-9. CROSSREF
5. Nuijten MB, Hartgerink CHJ, van Assen MALM, Epskamp S, Wicherts JM. The prevalence of statistical reporting errors in psychology (1985–2013). *Behav Res Methods* 2016;48(4):1205-26. PUBMED | CROSSREF
6. Jung RG, Di Santo P, Clifford C, Prosperi-Porta G, Skanes S, Hung A, et al. Methodological quality of COVID-19 clinical research. *Nat Commun* 2021;12(1):943. PUBMED | CROSSREF
7. Yoo JH. Defining the boundaries of AI use in scientific writing: a comparative review of editorial policies. *J Korean Med Sci* 2025;40(23):e187. PUBMED | CROSSREF
8. Doskaliuk B, Zimba O. Beyond the keyboard: academic writing in the era of ChatGPT. *J Korean Med Sci* 2023;38(26):e207. PUBMED | CROSSREF
9. Khalifa M, Albadawy M. Using artificial intelligence in academic writing and research: an essential productivity tool. *Comput Methods Programs Biomed Update* 2024;5:100145. CROSSREF
10. Brown NJL, Heathers JAJ. The GRIM Test: a simple technique detects numerous anomalies in the reporting of results in psychology. *Soc Psychol Personal Sci* 2017;8(4):363-9. CROSSREF
11. Nuijten MB, Polanin JR. *"statcheck"*: Automatically detect statistical reporting inconsistencies to increase reproducibility of meta-analyses. *Res Synth Methods* 2020;11(5):574-9. PUBMED | CROSSREF
12. Alnaimat F, AlSamhori ARF, Hamdan O, Seiil B, Qumar AB. Perspectives of artificial intelligence use for in-house ethics checks of journal submissions. *J Korean Med Sci* 2025;40(21):e170. PUBMED | CROSSREF
13. Kaur P, Stoltzfus J, Type I. II, and III statistical errors: a brief overview. *Int J Acad Med* 2017;3(2):268. CROSSREF
14. Thiese MS, Arnold ZC, Walker SD. The misuse and abuse of statistics in biomedical research. *Biochem Med (Zagreb)* 2015;25(1):5-11. PUBMED | CROSSREF
15. Gore A, Kadam Y, Chavan P, Dhumale G. Application of biostatistics in research by teaching faculty and final-year postgraduate students in colleges of modern medicine: a cross-sectional study. *Int J Appl Basic Med Res* 2012;2(1):11-6. PUBMED | CROSSREF
16. Flom P, Harron K, Ballesteros J, Kalinda C, Koutoumanou E, Miles J, et al. Common errors in statistics and methods. *BMJ Paediatr Open* 2024;8(1):e002755. PUBMED | CROSSREF
17. Bhathena RK. The 1995 pill scare and its aftermath: lessons learnt. *J Obstet Gynaecol* 1998;18(3):215-7. PUBMED | CROSSREF
18. Simkus A, Coolen-Maturi T, Coolen FPA, Bendtsen C. Statistical perspectives on reproducibility: definitions and challenges. *J Stat Theory Pract* 2025;19(3):40. CROSSREF
19. Allison DB, Brown AW, George BJ, Kaiser KA. Reproducibility: a tragedy of errors. *Nature* 2016;530(7588):27-9. PUBMED | CROSSREF
20. Calonge DS, Smail L, Kamalov F. Enough of the chit-chat: a comparative analysis of four AI chatbots for calculus and statistics. *J Appl Learn Teach* 2023;6(2). CROSSREF
21. Shumailov I, Shumaylov Z, Zhao Y, Papernot N, Anderson R, Gal Y. AI models collapse when trained on recursively generated data. *Nature* 2024;631(8022):755-9. PUBMED | CROSSREF
22. Wetzelhütter D, Prandner D. AI-enabled data analysis quality: addressing a knowledge gap. https://rgdoi.net/10.13140/RG.2.2.20359.37287. Updated 2023. Accessed October 3, 2025.
23. Shahrul AI, Syed Mohamed AMF. A comparative evaluation of statistical product and service solutions (SPSS) and ChatGPT-4 in statistical analyses. *Cureus* 2024;16(10):e72581. PUBMED | CROSSREF
24. Huang Y, Wu R, He J, Xiang Y. Evaluating ChatGPT-4.0's data analytic proficiency in epidemiological studies: a comparative analysis with SAS, SPSS, and R. *J Glob Health* 2024;14:04070. PUBMED | CROSSREF
25. Gibney E. AI tools are spotting errors in research papers: inside a growing movement. *Nature*. Forthcoming 2025. DOI: 10.1038/d41586-025-00648-5. PUBMED | CROSSREF
26. Pellegrina D, Helmy M. AI for scientific integrity: detecting ethical breaches, errors, and misconduct in manuscripts. *Front Artif Intell* 2025;8:1644098. PUBMED | CROSSREF
27. Takita H, Kabata D, Walston SL, Tatekawa H, Saito K, Tsujimoto Y, et al. A systematic review and meta-analysis of diagnostic performance comparison between generative AI and physicians. *NPJ Digit Med* 2025;8(1):175. PUBMED | CROSSREF
28. Chan J, Nsumba S, Wortsman M, Dave A, Schmidt L, Gollakota S, et al. Detecting clinical medication errors with AI enabled wearable cameras. *NPJ Digit Med* 2024;7(1):287. PUBMED | CROSSREF
29. Alqaraleh M, Almagharbeh WT, Ahmad MW. Exploring the impact of artificial intelligence integration on medication error reduction: a nursing perspective. *Nurse Educ Pract* 2025;86:104438. PUBMED | CROSSREF

30. Rada R. A case study of a retracted systematic review on interactive health communication applications: impact on media, scientists, and patients. *J Med Internet Res* 2005;7(2):e18. PUBMED | CROSSREF
31. Wilkinson J, Heal C, Antoniou GA, Flemyng E, Alfirevic Z, Avenell A, et al. Protocol for the development of a tool (INSPECT-SR) to identify problematic randomised controlled trials in systematic reviews of health interventions. *BMJ Open* 2024;14(3):e084164. PUBMED | CROSSREF
32. Abiola B, Pum M. A comparative study of Python and R for data science and statistical analysis. *Int J Nov Res Dev* 2024;8(11):719-23.
33. Nguyen MH, Vuong QH. Artificial intelligence and retracted science. *AI Soc* 2025;40(4):2345-6. CROSSREF
34. Campos-Varela I, Ruano-Raviña A. Misconduct as the main cause for retraction. A descriptive study of retracted publications and their authors. *Gac Sanit* 2019;33(4):356-60. PUBMED | CROSSREF
35. Hu W, Yan G, Zhang J, Chen Z, Qian Q, Wu S. Analysis of scientific paper retractions due to data problems: revealing challenges and countermeasures in data management. *Account Res*. Forthcoming 2025. DOI: 10.1080/08989621.2025.2531987. PUBMED | CROSSREF
36. Mehra MR, Ruschitzka F, Patel AN. Retraction-hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. *Lancet* 2020;395(10240):1820. PUBMED | CROSSREF
37. Fletcher AHA, Stevenson M. Predicting retracted research: a dataset and machine learning approaches. *Res Integr Peer Rev* 2025;10(1):9. PUBMED | CROSSREF
38. Doskaliuk B, Zimba O, Yessirkepov M, Klishch I, Yatsyshyn R. Artificial intelligence in peer review: enhancing efficiency while preserving integrity. *J Korean Med Sci* 2025;40(7):e92. PUBMED | CROSSREF
39. Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* 2000;16(5):412-24. PUBMED | CROSSREF
40. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. *ACM Trans Inf Syst* 2025;43(2):42:1-42. CROSSREF
41. Malik FS, Terzidis O. A hybrid framework for creating artificial intelligence-augmented systematic literature reviews. *Manage Rev Q*. Forthcoming 2025. DOI: 10.1007/s11301-025-00522-8. CROSSREF
42. Flanagin A, Kendall-Taylor J, Bibbins-Domingo K. Guidance for authors, peer reviewers, and editors on use of AI, language models, and chatbots. *JAMA* 2023;330(8):702-3. PUBMED | CROSSREF
43. Zhai C, Wibowo S, Li LD. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. *Smart Learn Environ* 2024;11(1):28. CROSSREF
44. Bakker CJ, Reardon EE, Brown SJ, Theis-Mahon N, Schroter S, Bouter L, et al. Identification of retracted publications and completeness of retraction notices in public health. *J Clin Epidemiol* 2024;173:111427. PUBMED | CROSSREF
45. Cao J, Yao J, Sun S, Song Z, Zhang F. Not all forms of artificial intelligence are perceived equal: AI functions and work outcomes. *J Open Innov Technol Mark Complex* 2025;11(2):100521. CROSSREF
46. Resnik DB, Hosseini M. Disclosing artificial intelligence use in scientific research and publication: When should disclosure be mandatory, optional, or unnecessary? *Account Res*. Forthcoming 2025. DOI: 10.1080/08989621.2025.2481949. PUBMED | CROSSREF
47. AlSamhori AF, Alnaimat F. Artificial intelligence in writing and research: ethical implications and best practices. *Cent Asian J Med Hypotheses Ethics* 2024;5(4):259-68. CROSSREF
48. Peh W, Saw A. Artificial intelligence: impact and challenges to authors, journals and medical publishing. *Malays Orthop J* 2023;17(3):1-4. PUBMED | CROSSREF
49. Lauer M, Wernimont A, Constant S. Using AI in peer review is a breach of confidentiality. https://www.csr.nih.gov/reviewmatters/2023/06/23/using-ai-in-peer-review-is-a-breach-of-confidentiality/. Updated 2023. Accessed October 4, 2025.
50. Hassija V, Chamola V, Mahapatra A, Singal A, Goel D, Huang K, et al. Interpreting black-box models: a review on explainable artificial intelligence. *Cognit Comput* 2024;16(1):45-74. CROSSREF
51. Kaebnick GE, Magnus DC, Kao A, Hosseini M, Resnik D, Dubljević V, et al. Editors' statement on the responsible use of generative AI technologies in scholarly journal publishing. *Med Health Care Philos* 2023;26(4):499-503. PUBMED | CROSSREF
52. Chen Z, Chen C, Yang G, He X, Chi X, Zeng Z, et al. Research integrity in the era of artificial intelligence: challenges and responses. *Medicine (Baltimore)* 2024;103(27):e38811. PUBMED | CROSSREF