

ARTICLE OPEN



Evaluation of ChatGPT-5-generated surgical literature: the accuracy of a review on penile prosthesis implantation

Fesih Ok¹, Ibrahim Halil Sukur¹, Murat Gül^{2,3✉} and Koenraad van Renterghem^{4,5,6}

© The Author(s) 2026

The integration of artificial intelligence language models into medical literature requires rigorous evaluation of accuracy and reliability, especially in specialized domains. This study assessed ChatGPT-5's capacity to generate clinically accurate scientific content on penile prosthesis implantation. Using structured prompts, ChatGPT-5 produced a narrative review evaluated across four domains: (1) verification of factual statements, (2) reference validity via PubMed and Google Scholar, (3) plagiarism screening with iThenticate and Quetext, and (4) qualitative assessment using Scale for the Assessment of Narrative Review Articles and a peer-review rubric. ChatGPT-5 demonstrated high factual accuracy overall, correctly supporting most statements, although errors were identified in historical timelines and survival data. In contrast, reference analysis revealed significant weaknesses, with only about one-third of citations being fully accurate and several containing fabricated or incomplete bibliographic details. Text similarity rates were low. Overall quality was rated as good according to standardized assessment tools, with strong agreement between reviewers. Collectively, these findings indicate that ChatGPT-5 can produce clinically accurate, well-structured content but demonstrates important weaknesses in reference reliability and evidence synthesis. The results support a hybrid model in which artificial intelligence serves as a drafting aid under expert supervision rather than as a standalone author. Future work should prioritize strengthening citation validity to enhance reliability while safeguarding scientific integrity.

IJIR: Your Sexual Medicine Journal; <https://doi.org/10.1038/s41443-026-01295-8>

INTRODUCTION

Artificial intelligence (AI)-based language models such as ChatGPT (OpenAI, San Francisco, CA, USA) have rapidly transformed the generation of scientific text, producing coherent and contextually relevant outputs in response to user prompts [1]. These tools are increasingly used in academic settings for literature summarization, drafting, and translation. However, their role in scientific authorship remains controversial, particularly following reports of ChatGPT being listed as a co-author and concerns regarding accountability and ethical standards [2].

Previous studies evaluating ChatGPT in medical contexts have identified important limitations, including factual inaccuracies, fabricated references, and insufficient domain-specific depth [3]. Although ChatGPT-5 incorporates advancements in language processing and contextual reasoning [4], its ability to generate scientifically reliable content in highly specialized surgical domains has not been systematically assessed. Given the potential risks of misinformation and the ethical implications for academic medicine, careful evaluation of AI-generated scientific texts is warranted before broader implementation [5].

Penile prosthesis implantation (PPI) represents a technically demanding and evidence-based surgical field, making it a suitable model for testing the reliability of AI-generated academic content [6]. The aim of this study was to evaluate the capacity of ChatGPT-5 to generate a narrative review on PPI and to systematically

assess its factual accuracy, reference validity, originality, and overall reporting quality.

MATERIALS AND METHODS

Study design and scope

This methodological evaluation study examined the ability of ChatGPT-5 to generate a narrative review on PPI. This procedure was selected as a test case because it represents a technically detailed and evidence-based urological procedure with an extensive and well-documented scientific literature, enabling structured comparison between AI-generated content and established knowledge.

As the study involved no human participants, clinical data, or identifiable health information, it did not require Institutional Review Board approval.

ChatGPT-5 narrative review generation

A narrative review on PPI was generated using ChatGPT-5 through a structured prompting strategy designed to reflect a standard academic use-case in which a researcher independently requests a narrative review from a large language model. The complete list of prompts is provided in the Supplementary Information (Supplementary File 1).

The process was initiated with the following instruction: "Write a draft of a scientific review on penile prosthesis implantation, covering clinical, surgical, and patient-centered aspects. We want our review to be concise, so limit your draft to the most relevant topics." In response, ChatGPT-5 generated an outline that defined the manuscript structure. The outline was not predefined by the authors but produced by the model itself.

¹Department of Urology, Adana City Training and Research Hospital, Adana, Turkey. ²Department of Urology, Faculty of Medicine Selçuk University, Konya, Turkey. ³Department of Andrology, Selçuk University School of Medicine, Konya, Turkey. ⁴Department of Urology, Jessa Hospital, Hasselt, Belgium. ⁵Department of Urology, Faculty of Medicine Hasselt University, Hasselt, Belgium. ⁶Department of Urology, University Hospital Leuven, Leuven, Belgium. ✉email: dmmuratgul@hotmail.com

Received: 4 November 2025 Revised: 6 May 2026 Accepted: 19 May 2026

Published online: 27 May 2026

Each section was subsequently generated individually using section-specific prompts derived from this AI-generated outline. For example, the surgical techniques section was generated using a targeted prompt specifying surgical approaches, management of complex anatomies, infection prevention, and perioperative care. After generating each section, ChatGPT-5 was instructed to provide five peer-reviewed references supporting the preceding content. The requirement to provide exactly five references per section was predefined to standardize output length across sections and to enable consistent comparative evaluation of citation accuracy.

The study was not designed to optimize prompt engineering performance or to compare alternative prompt formulations. No iterative refinement or performance-based prompt selection was performed. To assess response variability, selected prompts were repeated in separate sessions; however, only the first complete output for each predefined prompt was included in the formal evaluation to maintain methodological consistency and avoid selective reporting.

Evaluation of factual accuracy

All factual statements within the AI-generated review that would typically require citation in a peer-reviewed manuscript were identified and extracted. Two independent andrologists (MG, FO) with clinical expertise in PPI evaluated these statements. Each statement was verified through targeted searches in PubMed and Google Scholar. Statements were classified as:

- Supported (consistent with existing literature),
- Partially supported (incomplete or contextually inconsistent),
- Unsupported (no credible supporting evidence identified).

This classification framework (supported, partially supported, unsupported) is consistent with previously published methodologies evaluating AI-generated scientific content [7] and has been widely used to assess factual reliability in recent large language model studies. A statement was considered supported only when concordant with the preponderance of contemporary peer-reviewed evidence or established clinical guidelines, rather than on the basis of a single isolated publication. When variability existed across studies (e.g., differing survival or complication rates), priority was given to systematic reviews, meta-analyses, large contemporary cohort studies, or guideline statements. If evidence was conflicting or context-dependent, the statement was categorized as partially supported. Final classification required consensus between the two reviewers (MG, FO).

Assessment of citation accuracy

For each section, the five references provided by ChatGPT-5 were independently verified using PubMed, Google Scholar, CrossRef, and, when necessary, the official journal websites. Reference metadata—including title, authorship, journal name, publication year, volume/issue, page range, and DOI—were systematically cross-checked against the original publication records. References were classified into three predefined categories:

Fully accurate. The cited article corresponded to a real, peer-reviewed publication, and all major bibliographic elements (title, authorship, journal, year, and DOI or equivalent identifier) were correctly reported.

Partially inaccurate. A real and identifiable peer-reviewed publication existed; however, one or more bibliographic elements (e.g., author list, publication year, volume/issue number, page range, or DOI) contained inaccuracies or inconsistencies. Minor metadata errors did not qualify as fabrication if the underlying publication could be clearly verified.

Fabricated. No corresponding publication could be identified in major academic databases or journal archives, and the citation details did not match any existing peer-reviewed article.

This reference classification approach (fully accurate, partially inaccurate, fabricated) is consistent with previously reported methodologies used to evaluate citation accuracy and bibliographic reliability in AI-generated scientific content [7]. All classifications were performed independently by two reviewers (MG, FO). Discrepancies were resolved through consensus following re-verification of the source material.

Text integrity and plagiarism screening

To evaluate the originality of the manuscript generated by ChatGPT-5, the full text was uploaded to two widely used academic plagiarism detection

tools: iThenticate [8] and Quetext [9]. Both platforms provided an overall similarity index, and any matched content was further examined to determine whether it constituted legitimate overlap (e.g., common terminology) or potential plagiarism. Each sentence flagged by either tool was reviewed individually, and its potential source—if identified—was recorded along with the similarity percentage.

Qualitative assessment of ChatGPT-5 text quality

The quality of content generated by ChatGPT-5 was systematically evaluated using two complementary assessment frameworks to ensure comprehensive evaluation from multiple perspectives.

Primary assessment. The narrative review quality was further evaluated using the Scale for the Assessment of Narrative Review Articles (SANRA) [10], which provides a structured framework specifically designed for narrative reviews. SANRA comprises six validated items: justification of the article importance, statement of aims, description of the literature search, referencing quality, scientific reasoning, and appropriate presentation of data. Each item was scored from 0 (inadequate) to 2 (adequate), yielding a total possible score of 0–12 points.

Secondary assessment. A standardized peer review rubric was employed, encompassing eight core domains commonly used in academic journal evaluation processes: scientific accuracy, field relevance, structural clarity, content completeness, reference quality, innovation, clinical utility, and writing quality. Each domain was scored using a 5-point Likert scale (1 = poor, 5 = excellent), with cumulative scores classified as excellent (36–40), good (32–35), acceptable (28–31), needs improvement (24–27), or poor (< 24).

Assessment protocol. Two independent reviewers (MG, FO) with expertise in urology and evidence-based medicine conducted both evaluations. Each reviewer completed the assessments independently, and any scoring discrepancies were resolved through structured consensus discussion. Both item-level and total scores were recorded for each assessment tool, providing a comprehensive quality evaluation framework.

Endpoints

The primary endpoint of the study was the factual accuracy of statements generated by ChatGPT-5. Secondary endpoints included citation accuracy, text originality, and overall narrative quality as assessed by standardized evaluation tools.

Statistical analysis

All data were recorded and analyzed using SPSS version 27 (IBM Corp., Armonk, NY, USA). Descriptive statistics were calculated. Inter-rater reliability was assessed using the intraclass correlation coefficient based on a two-way mixed-effects model for absolute agreement. Reliability thresholds were interpreted according to established criteria.

RESULTS

AI-generated output and structure

ChatGPT-5 responded to all prompts within seconds and generated a structured narrative review on PPI, provided in Supplementary File 2. In response to the initial instruction, the model independently produced a seven-section outline encompassing introduction, clinical considerations, surgical techniques, outcomes and complications, patient-centered aspects, future directions, and conclusion (Fig. 1).

Repeated submission of identical prompts generated varying textual outputs (Fig. 2), indicating output variability across sessions despite unchanged instructions.

Title and keywords

ChatGPT-5 generated five alternative titles (Fig. 3), all aligned with the topic of PPI. The suggested keywords (“Penile prosthesis,” “Erectile dysfunction,” “Inflatable penile implant,” “Patient-centered care,” and “ChatGPT”) were considered relevant and reflective of the manuscript’s scope.

FO: Write the outline for a scientific review on penile prosthesis implantation, covering clinical, surgical, and patient-centered aspects. We want our review to be concise, so limit your draft to the most relevant topics.

ChatGPT: Title: A concise review on penile prosthesis implantation

1. Introduction

- Epidemiology of erectile dysfunction (ED) and indications for prosthesis
- Role of penile prosthesis compared to medical/conservative therapies
- Scope and aim of the review

2. Clinical Considerations

- Patient selection and indications
 - Vasculogenic, neurogenic, postsurgical, Peyronie's disease, refractory ED
- Preoperative assessment
 - Comorbidities (diabetes, cardiovascular, spinal cord injury)
 - Psychological evaluation and patient expectations
- Device options and evolution
 - Malleable vs inflatable prostheses
 - Technological advances (infection-resistant coatings, mechanical durability)

3. Surgical Techniques

- Surgical approaches: penoscrotal vs infrapubic (advantages, limitations)
- Technical pearls in difficult anatomy
 - Fibrosis, Peyronie's disease, revision surgery
- Infection prevention strategies
 - Antibiotic prophylaxis, "no-touch" technique, coated devices
- Perioperative management
 - Pain control, anticoagulation, catheterization protocols

4. Outcomes and Complications

- Mechanical reliability and device survival
- Patient and partner satisfaction
- Complication spectrum
 - Infection, erosion, mechanical failure, revision rates

5. Patient-Centered Aspects

- Counseling and shared decision-making
- Addressing psychosocial impact, body image, and partner perspectives
- Rehabilitation and sexual function beyond erection (orgasm, intimacy)

6. Future Directions

- Emerging device technologies (smart prostheses, minimally invasive approaches)
- Strategies to further reduce infection and mechanical failure
- Integration of regenerative medicine and prosthesis use

7. Conclusion

- Summary of key points
- Emphasis on individualized, multidisciplinary, patient-centered approach

Fig. 1 Response by ChatGPT-5 following a prompt to provide an outline for an article on penile prosthesis implantation. A response to the prompt: "Write the outline for a scientific review on penile prosthesis implantation, covering clinical, surgical, and patient-centered aspects. We want our review to be concise, so limit your draft to the most relevant topics".

Factual accuracy assessment

The AI-generated manuscript did not include in-text citations. Expert review identified 29 statements requiring academic references (Supplementary Table 1). Of these, three statements (10.3%) contained inaccuracies or internal inconsistencies.

Identified issues included:

- 1) An imprecise historical description regarding the introduction timeline of inflatable penile prostheses [11].
- 2) Conflicting reporting of 10-year device survival rates within different sections of the manuscript [12].
- 3) Inconsistent infection rate estimates across sections, reflecting variability in reported outcomes [13].

These discrepancies reflected internal inconsistency and contextual variability rather than pervasive factual unreliability. All remaining statements were verified and supported through targeted literature review.

Citation verification

ChatGPT-5 generated five references per section, resulting in a total of 28 citations.

Although most citation titles corresponded to real publications identifiable in academic databases, only 35.7% were fully accurate across all bibliographic parameters. Approximately 17.9% of references were entirely fabricated, while the remainder contained partial inaccuracies such as incorrect authorship details or mismatched publication metadata.

These findings indicate that while citation titles frequently appear plausible, bibliographic precision remains unreliable without independent verification. A detailed breakdown of citation validation outcomes is provided in the Supplementary File 1.

Text similarity and originality

Plagiarism screening demonstrated minimal similarity, with overall similarity indices of 3% (iThenticate) and 4% (Quetext). Review of flagged segments confirmed that overlaps consisted exclusively of common academic phrasing and standard scientific terminology.

FO: Write the 4th part of our scientific review (200 words) about outcomes and complications. The section should include: mechanical reliability and device survival, patient and partner satisfaction, and the complication spectrum.

ChatGPT: Modern three-piece inflatable penile prostheses (IPPs) demonstrate excellent mechanical reliability and long-term survival. Recent systematic review data report device survival rates of approximately 87% at 5 years, 77% at 10 years, and about 53% at 20 years, highlighting the durability of contemporary designs. Device longevity is influenced by patient factors, surgical technique, and appropriate prosthesis selection, underscoring the importance of individualized planning.

Patient and partner satisfaction remains consistently high, with both groups reporting improved sexual function and quality of life after implantation. For example, studies report patient satisfaction exceeding 85–90% and strong alignment with partner satisfaction. Satisfaction correlates strongly with preoperative counseling, expectation management, and uncomplicated postoperative recovery.

Despite advances, complications remain a critical consideration. Infection poses the most serious risk, affecting around 1–3% of primary implants and higher in revision cases, often necessitating explantation. Mechanical malfunctions, erosion, and cylinder migration may occur, particularly in patients with prior pelvic surgery or corporal fibrosis. Other risks include hematoma, glans necrosis, and urethral injury, though careful surgical technique, perioperative optimization, and adherence to infection-prevention protocols reduce their incidence. Comprehensive understanding of these outcomes enables clinicians to optimize results, guide patient expectations, and implement effective preventive strategies.

FO: Write the 4th part of our scientific review (200 words) about outcomes and complications. The section should include: mechanical reliability and device survival, patient and partner satisfaction, and the complication spectrum.

ChatGPT: Penile prosthesis implantation (PPI) is associated with high mechanical reliability and long-term device survival. Modern inflatable penile prostheses (IPPs), particularly those from leading manufacturers, demonstrate mechanical survival rates exceeding 90% at 10 years. Mechanical failure, though rare, is typically due to pump malfunction or cylinder leakage and often occurs years after implantation.

Patient and partner satisfaction rates are consistently high, ranging from 75% to over 90% in most studies. Satisfaction is primarily influenced by device performance, ease of use, postoperative counseling, and realistic expectations. Partners often report satisfaction levels similar to those of patients, particularly when involved in preoperative discussions.

Despite its efficacy, PPI carries a distinct spectrum of complications. The most concerning is infection, with an incidence of 1–3% in primary implants and up to 10% in revision surgeries. Biofilm-resistant coatings and improved surgical protocols have reduced infection rates significantly. Other complications include erosion, migration, hematoma formation, and mechanical failure. Corporal fibrosis, especially in revision cases or in patients with a history of priapism or Peyronie's disease, can increase technical difficulty and risk of complications. Despite these risks, PPI remains a reliable, durable, and highly satisfactory option for patients with medication-refractory erectile dysfunction.

Fig. 2 Variations in ChatGPT responses to an identical prompt at different times.

No evidence of substantive textual appropriation was identified (Table 1).

Qualitative quality assessment

Using the SANRA instrument, the AI-generated review received scores of 8 and 7 out of 12 from the two independent reviewers, indicating moderate-to-good narrative quality.

Assessment using the structured peer-review rubric yielded total scores of 33 and 32 out of 40, corresponding to a “Good” quality classification. Structural clarity and clinical relevance were rated strongly, whereas reference reliability represented the principal limitation.

Inter-rater reliability analysis demonstrated excellent agreement for both assessment tools (SANRA ICC = 0.95, 95% CI: 0.87–0.98; rubric ICC = 0.97, 95% CI: 0.91–0.99), confirming scoring consistency between reviewers.

DISCUSSION

This study provides a structured evaluation of ChatGPT-5's ability to generate clinically relevant and academically structured content in a specialized surgical domain. Our findings demonstrate that while the model produces coherent, well-organized, and largely accurate medical narratives, significant limitations remain in evidence synthesis and bibliographic reliability.

The overall factual consistency observed in this study is comparable to prior evaluations of earlier ChatGPT versions in medical contexts [7], suggesting incremental improvement in knowledge representation. However, the presence of internal inconsistencies—particularly regarding historical details, long-term device survival, and infection rates—highlights a persistent limitation: large language models (LLMs) do not inherently

reconcile conflicting evidence but instead reproduce patterns derived from heterogeneous training data. This inability to critically synthesize discordant findings has direct clinical implications, particularly in areas such as patient counseling and shared decision-making, where precise risk communication is essential [14, 15].

The most substantial limitation identified in our analysis concerns reference reliability. Fabricated and partially inaccurate citations remain a fundamental challenge across ChatGPT iterations [7, 16]. The phenomenon of “hallucinated” references is particularly problematic because fabricated citations are often structurally plausible, occasionally incorporating legitimate author names or realistic bibliographic formatting. Such sophistication increases the risk of undetected misinformation and may compromise academic integrity [17].

This pattern is consistent with broader investigations of AI-generated content. Large-scale media evaluations have demonstrated frequent source attribution errors, inaccurate claims, and insufficient contextualization across multiple chatbot systems [18]. The persistence of bibliographic distortion suggests that current language models simulate citation structure rather than retrieve verified sources. Consequently, traditional editorial screening may be insufficient, and structured verification protocols are necessary when AI-assisted writing is employed [19].

In contrast, the plagiarism analysis yielded reassuring findings. The minimal similarity indices and absence of substantive textual overlap align with previous research indicating that AI-generated text generally produces original surface-level phrasing while conforming to academic conventions [20]. However, originality at the textual level does not equate to epistemic reliability, and the distinction between linguistic originality and factual validity warrants continued scholarly attention. Moreover, conventional

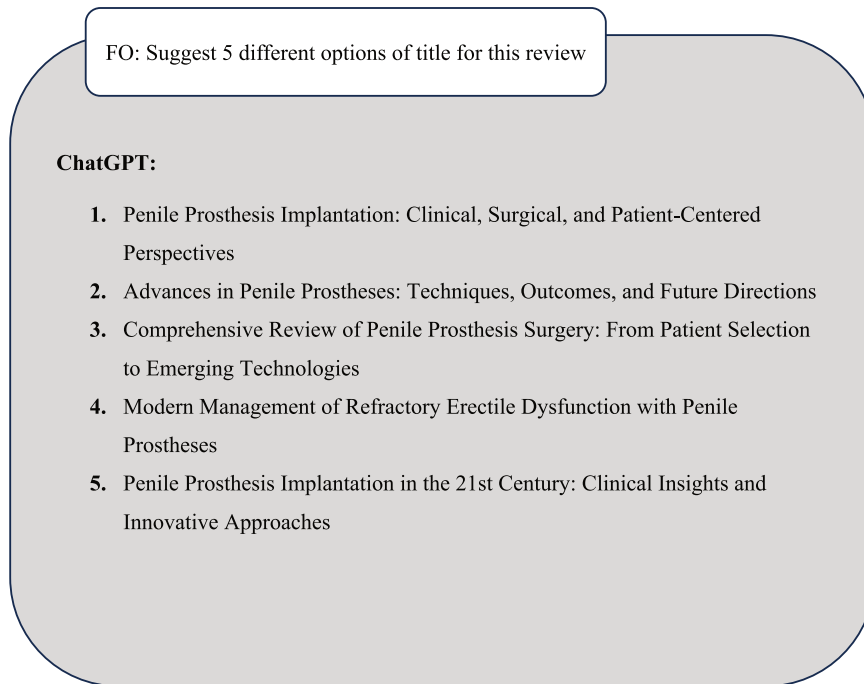


Fig. 3 A ChatGPT-5 response enumerating five potential titles for the review.

Table 1. Text similarity analysis reports.

Plagiarism checker	% of plagiarism	Possible plagiarized sentences	Reference title
Ithenticate	3	"for men with erectile dysfunction (ED) refractory to medical therapy."	Penile Prosthesis Complications: Planning, Prevention, and Decision Making [24]
		"Shared decision-making with the patient and often"	Management Reasoning: Beyond the Diagnosis [25]
Quetext	4	"Penile prosthesis implantation remains the definitive surgical solution for men with erectile dysfunction (ED) refractory to medical therapy"	Narrative review of immediate salvage for penile prosthesis infection [26]
		"Body image, self-esteem, and sexual identity may be"	Cancer and Sexual Health https://link.springer.com/book/10.1007/978-1-60761-916-1
		"sexual health and quality of life"	Other (Bladder Cancer, Sexual Dysfunction) https://link.springer.com/chapter/10.1007/978-3-319-63910-9_15

plagiarism detection tools rely on surface-level string-matching algorithms and may fail to identify statistically derivative or semantically reconstructed content generated by LLMs trained on vast scientific corpora. Thus, even in the absence of overt textual duplication, AI-generated writing may reflect probabilistic recombination of existing knowledge rather than true conceptual independence. Evaluating deeper intellectual originality in AI-assisted scientific writing remains an unresolved methodological challenge.

Quality assessment results indicate that ChatGPT-5 can generate structurally coherent and clinically relevant drafts that meet moderate-to-good narrative standards. Nevertheless, deficiencies in methodological transparency, literature synthesis depth, and referencing rigor remain evident [21]. These weaknesses underscore the distinction between narrative fluency and scientific rigor.

The observed variability in outputs generated from identical prompts raises additional concerns regarding reproducibility. While variability may enhance creative flexibility, it conflicts with the reproducibility principles fundamental to scientific research.

Inconsistent outputs may introduce unintended bias or incompleteness in AI-assisted academic writing [22].

Future investigations should move beyond single-session evaluations and systematically examine output variability across multiple independent generations. Multi-run study designs incorporating repeated prompting under standardized conditions would enable characterization of performance distributions, stability metrics, and dispersion of quality scores. Such approaches would provide a more rigorous assessment of model robustness and reproducibility in scientific writing contexts.

From a practical perspective, our findings support a supervised implementation model. ChatGPT-5 demonstrates potential utility as an educational drafting tool, particularly for generating structured overviews of complex surgical procedures. However, expert oversight is essential to verify factual accuracy, reconcile conflicting evidence, and validate all citations prior to academic or clinical use.

Ethically, these findings argue against the use of large LLMs as autonomous authors in medical literature. Instead, a collaborative

framework—where AI assists in drafting while human experts retain responsibility for validation and interpretation—appears most appropriate. Future advancements should prioritize integration of real-time source verification systems and improved evidence-tracing mechanisms to reduce citation fabrication and enhance epistemic reliability [23].

Several limitations of this study merit consideration. The evaluation focused on a single surgical domain, and findings may not generalize across specialties. Accuracy assessment relied on expert appraisal, which inherently involves subjective judgment despite structured evaluation. Inter-rater reliability estimates were derived from two reviewers evaluating a single AI-generated document and may overestimate agreement compared to evaluations involving multiple documents or a larger and more heterogeneous panel of raters. The prompts were developed by a single author and alternative prompt formulations were not systematically tested; therefore, different prompting strategies may have influenced output quality. The analysis reflects performance based on a single-session output per prompt and does not capture potential variability across repeated generations. The study was not designed to characterize distributional performance across multiple independent generations; therefore, variability in quality metrics and dispersion measures across outputs were not assessed. In addition, the study was conducted using the publicly available web-based ChatGPT interface without access to a fixed API version or model snapshot. Given that LLMs are updated dynamically and without publicly accessible version control, exact reproducibility of outputs cannot be guaranteed even with identical prompts. Furthermore, due to the rapid evolution of AI systems, performance characteristics may change over time. Broader, multi-domain, and longitudinal investigations are warranted to further define the role of AI in medical scientific writing.

CONCLUSION

In this evaluation of a ChatGPT-5-generated narrative review on PPI, the model produced structured and largely accurate clinical content but demonstrated important limitations in reference reliability and internal consistency. Given that the analysis was restricted to a single surgical topic, the findings cannot be generalized to broader scientific writing. ChatGPT-5 may serve as a drafting aid under expert supervision; however, citation verification and critical appraisal remain essential before academic or clinical use.

DATA AVAILABILITY

Additional data are available from the corresponding author on reasonable request.

REFERENCES

1. Storey VC, Yue WT, Zhao JL, Lukyanenko R. Generative artificial intelligence: evolving technology, growing societal impact, and opportunities for information systems research. *Inf Syst Front*. 2025. <https://doi.org/10.1007/s10796-025-10581-7>.
2. Perkins M, Roe J. Academic publisher guidelines on AI usage: a ChatGPT supported thematic analysis. *F1000Res*. 2024;12:1398 <https://doi.org/10.12688/f1000research.142411.2>.
3. Bhattacharyya M, Miller VM, Bhattacharyya D, Miller LE. High rates of fabricated and inaccurate references in ChatGPT-generated medical content. *Cureus*. 2023;15:e39238 <https://doi.org/10.7759/cureus.39238>.
4. Daulat S, Dholaria N, Burnet G, Patil S, Manne B, Choudhary A, et al. Prompt engineering and follow-up questioning improves the readability of spine surgery questions in large language models. *World Neurosurg*. 2025;203:124423 <https://doi.org/10.1016/j.wneu.2025.124423>.
5. Chustecki M. Benefits and risks of AI in health care: narrative review. *Interact J Med Res*. 2024;13:e53616 <https://doi.org/10.2196/53616>.

6. Ioannis M, Georgios L, Stavros T, Anastasios A, Koenraad VR. Update on penile prosthesis. *Asian J Androl*. 2025;27:1–8. <https://doi.org/10.4103/aja2024100>.
7. Safrai M, Orwig KE. Utilizing artificial intelligence in academic writing: an in-depth evaluation of a scientific review on fertility preservation written by ChatGPT-4. *J Assist Reprod Genet*. 2024;41:1871–80. <https://doi.org/10.1007/s10815-024-03089-7>.
8. iThenticate. 2025; Available from: <https://www.ithenticate.com/>.
9. Quetext. 2025; Available from: <https://www.quetext.com/>.
10. Baethge C, Goldbeck-Wood S, Mertens S. SANRA—a scale for the quality assessment of narrative review articles. *Res Integr Peer Rev*. 2019;4:5. <https://doi.org/10.1186/s41073-019-0064-8>.
11. Rodriguez KM, Pastuszak AW. A history of penile implants. *Transl Androl Urol*. 2017;6:851–7. <https://doi.org/10.21037/tau.2017.04.02>.
12. Bellaiche J, Gonzalez Serrano A, Larre S. Survie et facteurs associés à la durée de vie des prothèses péniennes à long terme [Survival and associated factors with long-term survival of penile prosthesis]. *Prog Urol*. 2021;31:374–80. <https://doi.org/10.1016/j.purol.2020.09.018>.
13. Eid JF, Wilson SK, Cleves M, Salem EA. Coated implants and “no touch” surgical technique decreases risk of infection in inflatable penile prosthesis implantation to 0.46%. *Urology*. 2012;79:1310–5. <https://doi.org/10.1016/j.urology.2011.11.076>.
14. Zybaczynska J, Norris M, Modi S, Brennan J, Jhaveri P, Craig TJ. Artificial intelligence-generated scientific literature: a critical appraisal. *J Allergy Clin Immunol Pr*. 2024;12:106–10. <https://doi.org/10.1016/j.jaip.2023.10.010>.
15. Peng Y, Rousseau JF, Shortliffe EH, Weng C. AI-generated text may have a role in evidence-based medicine. *Nat Med*. 2023;29:1593–4. <https://doi.org/10.1038/s41591-023-02366-9>.
16. Chelli M, Descamps J, Lavoué V, Trojani C, Azar M, Deckert M, et al. Hallucination rates and reference accuracy of ChatGPT and bard for systematic reviews: comparative analysis. *J Med Internet Res*. 2024;26:e53164 <https://doi.org/10.2196/53164>.
17. Jain A, Nimonkar P, Jadhav P. Citation integrity in the age of AI: evaluating the risks of reference hallucination in maxillofacial literature. *J Craniomaxillofac Surg*. 2025;53:1871–2. <https://doi.org/10.1016/j.jcms.2025.08.004>.
18. Nicolay D, Vandendriessche B. AI-chatbots vaak in de fout bij vragen over nieuws en actualiteit: problemen in bijna helft antwoorden. 2025. <https://www.vrt.be/vrtnws/nl/2025/10/20/onderzoek-ai-chatbots/>.
19. Elali FR, Rachid LN. AI-generated research paper fabrication and plagiarism in the scientific community. *Patterns (NY)*. 2023;4:100706 <https://doi.org/10.1016/j.patter.2023.100706>.
20. Jairoon AA, El-Dahiyat F, ElRefae GA, Al-Hemyari SS, Shahwan M, Zyoud SH, et al. Detecting manuscripts written by generative AI and AI-assisted technologies in the field of pharmacy practice. *J Pharm Policy Pract*. 2024;17:2303759. <https://doi.org/10.1080/20523211.2024.2303759>.
21. Iqbal U, Tanweer A, Rahmanti AR, Greenfield D, Lee LT, Li YJ. Impact of large language model (ChatGPT) in healthcare: an umbrella review and evidence synthesis. *J Biomed Sci*. 2025;32:45 <https://doi.org/10.1186/s12929-025-01131-z>.
22. Lau J, Bisht S, Horton R, Crisan A, Jones J, Gantotti S, et al. Creation of scientific response documents for addressing product medical information inquiries: mixed method approach using artificial intelligence. *JMIR AI*. 2025;4:e55277 <https://doi.org/10.2196/55277>.
23. Kacena MA, Plotkin LI, Fehrenbacher JC. The use of artificial intelligence in writing scientific review articles. *Curr Osteoporos Rep*. 2024;22:115–21. <https://doi.org/10.1007/s11914-023-00852-0>.
24. Scherzer ND, Dick B, Gabrielson AT, Alzweri LM, Hellstrom WJG. Penile prosthesis complications: planning, prevention, and decision making. *Sex Med Rev*. 2019;7:349–59. <https://doi.org/10.1016/j.sxmr.2018.04.002>.
25. Cook DA, Sherbino J, Durning SJ. Management reasoning: beyond the diagnosis. *JAMA*. 2018;319:2267–8. <https://doi.org/10.1001/jama.2018.4385>.
26. Adams ES, Tua-Caraccia RD, Lentz AC. Narrative review of immediate salvage for penile prosthesis infection. *Transl Androl Urol*. 2024;13:584–95. <https://doi.org/10.21037/tau-23-277>.

AUTHOR CONTRIBUTIONS

FO: the conception and design of the study, acquisition of data, analysis and interpretation of data, literature search, drafting the article. MG: critical revision of the manuscript for important intellectual content, final approval of the version to be submitted. IHS: the conception and design of the study, acquisition of data, analysis, and interpretation of data, literature search. KVR: critical revision of the manuscript for important intellectual content

FUNDING

Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK).

COMPETING INTERESTS

The authors declare no competing interests.

ETHICS APPROVAL

This article does not contain any studies with human participants or animals performed by any of the authors. Therefore, there is no need for ethical approval.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41443-026-01295-8>.

Correspondence and requests for materials should be addressed to Murat Gül.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2026