ORIGINAL RESEARCH ORİJİNAL ARAŞTIRMA

DOI: 10.5336/urology.2025-109438

Comparison of Artificial Intelligence Responses About Penile Prosthesis implantation in Terms of Quality and Readability: A Methodological and Validity Study on ChatGPT, Gemini, and DeepSeek

Penil Protez İmplantasyonu Hakkında Yapay Zeka Yanıtlarının Kalite ve Okunabilirlik Açısından Karşılaştırılması: ChatGPT, Gemini ve DeepSeek Üzerine Metodolojik ve Geçerlilik Çalışması

Murat DEMİR[a], Nedim BEDİR[a], Doğan BOĞA[a], Recep ERYILMAZ[a], Rahmi ASLAN[a], Kasım ERTAŞ[b], Kerem TAKEN[a]

[a]Van Yüzüncü Yıl University Faculty of Medicine, Department of Urology, Van, Türkiye
[b]Lokman Hekim Van Hospital, Clinic of Urology, Van, Türkiye

ABSTRACT Objective: The use of artificial intelligence models to inform patients about penile prosthesis implantation has not been sufficiently studied in the literature. This study evaluates the quality and readability of responses given by ChatGPT, DeepSeek, and Gemini to the most common questions asked by patients before penile prosthesis implantation. **Material and Methods:** We selected 15 common patient questions related to preoperative management, complications, and postoperative care regarding penile prosthesis implantation. These questions were submitted in English on February 10, 2025, to ChatGPT, DeepSeek, and Gemini. Responses were analyzed for information quality (using the DISCERN Tool) and readability (using the Gunning Fog Index). **Results:** ChatGPT had the highest DISCERN score (4.2±0.3), while DeepSeek (3.8±0.4) and Gemini (3.5±0.5) had lower scores ($p<0.05$ for all comparisons). In terms of readability, ChatGPT generated the most comprehensible responses (Gunning Fog Index: 7.1±0.6), whereas DeepSeek (8.3±0.7) and Gemini (9.0±0.8) produced more complex texts ($p<0.05$ for all comparisons). **Conclusion:** ChatGPT outperformed the other models in terms of information quality and readability, demonstrating its potential as a complementary tool for patient education and informed consent. However, the observed differences in information accuracy and readability among the models highlight the need for further refinement before full integration into clinical use.

Keywords: Intelligence; artificial; penile prosthesis implantation

ÖZET Amaç: Yapay zeka modellerinin hastaları penil protez implantasyonu hakkında bilgilendirme amacıyla kullanımı literatürde yeterince incelenmemiştir. Bu çalışma, hastalar tarafından penil protez implantasyonu öncesinde en sık sorulan sorulara ChatGPT, DeepSeek ve Gemini tarafından verilen yanıtların kalite ve okunabilirlik açısından değerlendirilmesini amaçlamaktadır. **Gereç ve Yöntemler: Penil pro**tez implantasyonu ile ilgili preoperatif yönetim, komplikasyonlar ve postoperatif bakım konularında en sık sorulan 15 hasta sorusu seçildi. Bu sorular, 10 Şubat 2025 tarihinde İngilizce olarak ChatGPT, DeepSeek ve Gemini'ye yönlendirildi. Yanıtlar, bilgi kalitesi açısından DISCERN Aracı kullanılarak ve okunabilirlik açısından Gunning Fog İndeksi ile analiz edildi. **Bulgular:** ChatGPT, en yüksek DISCERN puanına sahipti (4,2±0,3), bunu DeepSeek (3,8±0,4) ve Gemini (3,5±0,5) takip etti (tüm karşılaştırmalarda $p<0,05$). Okunabilirlik açısından ise ChatGPT en anlaşılır yanıtları üretti (Gunning Fog İndeksi: 7,1±0,6), buna karşın DeepSeek (8,3±0,7) ve Gemini (9,0±0,8) daha karmaşık metinler oluşturdu (tüm karşılaştırmalarda $p<0,05$). **Sonuç:** ChatGPT, bilgi kalitesi ve okunabilirlik açısından diğer modellere üstünlük sağlamış ve hasta eğitimi ile bilgilendirilmiş onam sürecinde tamamlayıcı bir araç olarak potansiyelini göstermiştir. Ancak, modeller arasındaki bilgi doğruluğu ve okunabilirlik farklılıkları, klinik kullanıma tam entegrasyon sağlanmadan önce daha fazla iyileştirme gerekliliğini ortaya koymaktadır.

Anahtar Kelimeler: Zekâ; yapay; penil protez implantasyonu

TO CITE THIS ARTICLE:

Demir M, Bedir N, Boğa D, Eryılmaz R, Aslan R, Ertaş K, et al. Comparison of artificial intelligence responses about penile prosthesis implantation in terms of quality and readability: A methodological and validity study on ChatGPT, Gemini, and DeepSeek. J Reconstr Urol. 2025;15(1):21-6.

Correspondence: Murat DEMİR
Van Yüzüncü Yıl University Faculty of Medicine, Department of Urology, Van, Türkiye
E-mail: urologmurat72@gmail.com

Peer review under responsibility of Journal of Reconstructive Urology.

Received: 12 Feb 2025 ***Received in revised form:*** 07 Mar 2025 ***Accepted:*** 07 Mar 2025 ***Available online***: 27 Mar 2025

Erectile dysfunction (ED) is a common health problem that significantly affects sexual health and reduces the quality of life in men.[1] In recent years, artificial intelligence (AI)-assisted language models have been increasingly used in areas such as access to medical information and patient education. Popular AI models such as ChatGPT (OpenAI, USA), DeepSeek (China), and Gemini (Google, USA) have gained attention for their ability to instantly respond to users' health-related questions.[2] However, the accuracy, reliability, and comprehensibility of the information provided by these models on medical issues-especially on a sensitive and complex subject such as erectile dysfunction-have not yet been comprehensively evaluated.

Informed consent forms the basis of ethical medical practice by ensuring that patients have adequate information about treatment options, potential risks, and postoperative expectations, particularly in elective surgery. However, due to the difficulty of understanding complex medical information, patients often experience dissatisfaction or decision regret. Natural language processing (NLP) models have the potential to improve surgeon-patient interaction and support the informed consent process by enhancing patients' understanding of procedures. In the literature, various applications of NLP models such as ChatGPT have been demonstrated, ranging from structuring radiology reports to supporting informed consent in orthopedic surgery and urology. However, inconsistencies have been reported regarding the accuracy and medical appropriateness of the information produced by these models.[3]

As a complementary tool to traditional face-to-face consultations, NLP models can help patients become more informed and prepared before undergoing a procedure. Additionally, after the consultation, these models can reinforce the information provided, thereby enhancing patient education and satisfaction. In this context, integrating NLP-based AI models into clinical practice may optimize patient-surgeon interaction and strengthen the concept of patient-centered care.

In this study, we will evaluate the information quality and readability of the answers provided by ChatGPT, DeepSeek, and Gemini to frequently asked patient questions during face-to-face consultations before penile prosthesis implantation. Additionally, we will examine the potential of these models as supportive tools in the informed consent process.

MATERIAL AND METHODS

This study was conducted using the 15 most frequently asked questions by patients before penile prosthesis implantation (PPI), which were previously used in the study by Schmidt et al. Of these questions, 5 were related to preoperative management before PPI, 5 were related to complications, and 5 were related to postoperative care. This study was conducted in accordance with the criteria of the Declaration of Helsinki. Ethics committee approval was not obtained for this study as no patient data were used.

DATA COLLECTION

AI Models

ChatGPT, DeepSeek, and Gemini were used in this study. Consistency was ensured by asking the same questions to each model. The questions were submitted in English to ChatGPT, DeepSeek, and Gemini via a web provider on February 10, 2025. Access to the models was provided through official application programming interfaces or web interfaces. The latest version of each model was used.

EVALUATION CRITERIA

The following criteria and metrics were used to assess the quality and clarity of the responses:

Information Quality (DISCERN Tool):[4,5] The quality of the information provided was evaluated using the DISCERN tool. DISCERN consists of a 16-item questionnaire that measures the reliability, objectivity, and comprehensiveness of health information. Each question was scored using a Likert scale from 1 to 5.

Some Examples

- Are the sources of information clearly indicated?

- Are treatment options presented in a balanced way?
- Is the information up-to-date and reliable?

Readability (Gunning Fog Index)[6]

The readability of the responses was measured using the Gunning Fog Index. This index evaluates text comprehensibility based on sentence length and the usage of complex words. The formula is as follows:

Gunning Fog Index=0.4×(Total Word Count: Number of Sentences+Percentage of Complex Words)

Interpretation:

- Lower index values (e.g., 6-8) indicate the readability level suitable for 6th to 8th grade students.
- Higher values (12 and above) indicate the readability level suitable for 12th grade and above students

STATISTICAL ANALYSIS

Statistical Methods: The significance of differences between the models' performances was evaluated using Analysis of variance and "post-hoc" tests to compare the DISCERN and Gunning Fog Index scores.

RESULT

INFORMATION QUALITY (DISCERN SCORES)

Overall Performance: The DISCERN scores of the models showed significant differences in terms of information quality. ChatGPT had the highest information quality, with an average score of 4.2±0.3, while DeepSeek scored 3.8±0.4 and Gemini scored 3.5±0.5 (Table 1).

OVERALL READABILITY (GUNNING SCORES)

ChatGPT provided the most comprehensible answers, with an average index value of 7.1±0.6. DeepSeek used a more complex language, with an index value of 8.3±0.7, and Gemini had the most complex responses, with an index value of 9.0±0.8 (Table 1).

Best Performance: ChatGPT demonstrated the best performance in terms of both information quality and readability. This model used a balanced and comprehensible language, making it easier for users to access medical information; DeepSeek: While it performed similarly to ChatGPT in terms of information quality, it lagged behind in readability due to its use of more technical language; Gemini: Compared to the other models, Gemini showed lower performance in both information quality and readability. Its weaknesses were particularly evident in its complex sentence structures and lack of source attribution.

DISCUSSION

Physicians often have a busy work schedule and, when recommending surgical intervention to patients, they must provide detailed information about the indication for the operation, the surgical technique to be performed, potential complications, and postoper-

TABLE 1: Comparison of different AIModels in terms of medical content quality and readability

Measurement	ChatGPT	DeepSeek	Gemini	p value
DISCERN Score	4.2±0.3	3.8±0.4	3.5±0.5	0.02 (ChatGPT vs DeepSeek) 0.04 (DeepSeek vs Gemini) 0.06 (ChatGPT vs Gemini)
Gunning Fog Index	7.1±0.6	8.3±0.7	9.0±0.8	0.01 (ChatGPT vs DeepSeek) 0.03 (DeepSeek vs Gemini) 0.05 (ChatGPT vs Gemini)

ative care. However, due to limited consultation time, patients often do not receive sufficient information or seek additional resources to supplement the information provided.[7] With the advancement of AI technologies, patients now have access to health-related information through AI-assisted systems. This study evaluates the accuracy and readability of surgical information provided by AI models (ChatGPT, DeepSeek, and Gemini).

This study aims to compare the information quality and readability of responses given by AI models (ChatGPT, DeepSeek, and Gemini) regarding ED. The findings reveal significant differences in these models' capacities to provide medical information and their user-friendly approaches.

In the study, it was found that ChatGPT had the highest information quality in terms of DICERN scores. This result can be explained by ChatGPT being trained on a large and up-to-date dataset. Additionally, ChatGPT's comprehensive explanations of treatment options and risk factors suggest that it could assist users in making informed decisions. Similarly, DeepSeek also demonstrated satisfactory performance in terms of information quality. However, the fact that Gemini provided incomplete or superficial information in some questions indicates that this model needs improvement in delivering medical information.

In the literature, the reliability of AI models in delivering medical information is a frequently debated topic. In particular, the recency and scope of the models' training data directly impact the quality of the information.[8] While ChatGPT can provide accurate and reliable information in specific contexts, it may fall short in terms of quality when dealing with complex or detailed topics. The accuracy of the model's responses is generally dependent on the reliability of the internet-based sources it draws upon. Previous studies have shown that health information derived from the internet can be inconsistent, and this inconsistency can also affect AI responses.[9]

Studies on the use of AI in urology have shown that the model is largely consistent with the European Urology Association guidelines in diagnosis and treatment planning; however, there are significant gaps in surgical planning and long-term follow-up. Furthermore, different studies have reported varying results regarding the quality of AI's informational output. For instance, some studies found the responses AI models give to patient inquiries to be largely satisfactory, while others indicated that the information quality was only moderate.[10] The findings of this study suggest that AI models hold promise in providing medical information, but there are still areas in need of improvement.

In terms of readability, ChatGPT's Gunning Fog Index scores were found to be lower compared to the other models. This indicates ChatGPT's success in conveying medical terminology in a simple and understandable manner. This feature is particularly valuable for users with limited medical knowledge and enhances the potential of AI models in the healthcare field. DeepSeek and Gemini, on the other hand, tended to use more technical language, which reduces accessibility for general users. In the literature, the readability of health information is emphasized as being critical for patient education and treatment adherence.[11] The findings of this study highlight the need for improvement in the readability of AI models, particularly in conveying complex medical topics in simple language.

The results of the study show significant performance differences between the models. ChatGPT's superior performance suggests that this model is more advanced than the others in terms of both information quality and user-friendly approach. However, the performances of DeepSeek and Gemini indicate that these models also have the potential for improvement. In particular, improving Gemini's information quality and readability could increase its use in the healthcare field. The limitations of AI models in providing medical information are frequently discussed in the literatüre.[12] The findings of this study support the idea that improvements are needed in the models' training data's recency, scope, and user-friendly approaches.

This study has some limitations. First, the study only evaluated questions related to penile prosthesis implantation. No similar evaluation was conducted for other medical topics. Secondly, the recency and scope of the models' training data can affect the quality of their answers. Finally, the metrics used in the study, such as DICERN and Gunning Fog, are qualitative assessments, meaning they may involve subjective judgments.

CONCLUSION

This study compared the information quality and readability of answers provided by AI models like ChatGPT, DeepSeek, and Gemini regarding ED. The findings show that ChatGPT outperforms the other models in terms of both information quality and user-friendly approach. However, it also became clear that all models have areas that need improvement. These findings lay an important foundation for the potential of AI models in healthcare and for future research.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and / or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

Idea/Concept: *Murat Demir, Nedim Bedir, Doğan Boğa;* ***Design:*** *Murat Demir, Nedim Bedir, Recep Eryılmaz;* ***Control/Supervision:*** *Murat Demir, Recep Eryılmaz, Rahmi Aslan;* ***Data Collection and/or Processing:*** *Murat Demir, Rahmi Aslan, Kasım Ertaş;* ***Analysis and/or Interpretation:*** *Murat Demir, Kasım Ertaş, Kerem Taken;* ***Literature Review:*** *Murat Demir, Kerem Taken, Nedim Bedir;* ***Writing the Article:*** *Murat Demir, Nedim Bedir, Doğan Boğa, Recep Eryılmaz, Rahmi Aslan, Kasım Ertaş, Kerem Taken;* ***Critical Review:*** *Murat Demir;* ***References and Fundings:*** *Murat Demir;* ***Materials:*** *Murat Demir.*

REFERENCES

1. Bajic P, Patel PM, Nelson MH, Dornbier RA, Kirshenbaum EJ, Baker MS, et al. Penile prosthesis implantation and timing disparities after radical prostatectomy: results from a statewide claims database. J Sex Med. 2020;17(6):1175-81. [Crossref] [PubMed]
2. Homolak J. Opportunities and risks of ChatGPT in medicine, science, and academic publishing: a modern Promethean dilemma. Croat Med J. 2023;64(1):1-3. [Crossref] [PubMed] [PMC]
3. Shayegh NA, Byer D, Griffiths Y, Coleman PW, Deane LA, Tonkin J. Assessing artificial intelligence responses to common patient questions regarding inflatable penile prostheses using a publicly available natural language processing tool (ChatGPT). Can J Urol. 2024;31(3):11880-5. [PubMed]
4. Schmidt J, Lichy I, Kurz T, Peters R, Hofbauer S, Plage H, et al. ChatGPT as a support tool for informed consent and preoperative patient education prior to penile prosthesis implantation. J Clin Med. 2024;13(24):7482. [Crossref] [PubMed] [PMC]
5. Rees CE, Ford JE, Sheard CE. Evaluating the reliability of DISCERN: a tool for assessing the quality of written patient information on treatment choices. Patient Educ Couns. 2002;47(3):273-5. [Crossref] [PubMed]
6. Świeczkowski D, Kułacz S. The use of the Gunning Fog Index to evaluate the readability of Polish and English drug leaflets in the context of Health Literacy challenges in Medical Linguistics: An exploratory study. Cardiol J. 2021;28(4):627-31. [Crossref] [PubMed] [PMC]
7. Deng J, Lin Y. The benefits and challenges of ChatGPT: An overview. Benefits. 2023;2(2):2022. [Crossref]
8. Smith A, Hachen S, Schleifer R, Bhugra D, Buadze A, Liebrenz M. Old dog, new tricks? Exploring the potential functionalities of ChatGPT in supporting educational methods in social psychiatry. Int J Soc Psychiatry. 2023;69(8): 1882-9. [Crossref] [PubMed]
9. Capece M, Di Giovanni A, Cirigliano L, Napolitano L, La Rocca R, Creta M, et al. YouTube as a source of information on penile prosthesis. Andrologia. 2022;54(1):e14246. [Crossref] [PubMed]
10. Talyshinskii A, Juliebø-Jones P, Zeeshan Hameed BM, Naik N, Adhikari K, Zhanbyrbekuly U, et al. ChatGPT as a clinical decision maker for urolithiasis: compliance with the current european association of urology guidelines. Eur Urol Open Sci. 2024;69:51-62. [Crossref] [PubMed] [PMC]
11. Biswas SS. Role of ChatGPT in public health. Ann Biomed Eng. 2023;51(5):868-9. [Crossref]
12. Mattas PS. ChatGPT: a study of AI language processing and its implications. International Journal of Research Publication and Reviews. 2023;4(2):435-40. [Crossref]