Original Article
Özgün Araştırma

Endourol Bull. 2026;18(1):30-38. doi: 10.54233/endourolbull-1769794

A Comparison Study of AI Response Quality in Kidney Stone-Related Questions: Google vs. ChatGPT

Böbrek Taşıyla İlgili Sorularda Yapay Zekâ (ChatGPT) ve Google Cevap Kalitesinin Karşılaştırılması

Burak Elmaağaç, Abdullah Gölbaşı, Ali Yasin Özercan, Hüseyin Biçer

Department of Urology, University of Health Sciences, Kayseri City Hospital, Kayseri, Türkiye

ABSTRACT

Objective: This study intends to evaluate, in Türkiye during the previous 12 months, the readability, reliability, and general quality of responses given by Google and ChatGPT on often-requested inquiries about kidney stones.

Material And Methods: The 10 most often searched queries and 4 trending topics on kidney stones were found by means of Google Trends data. Fourteen queries in all were asked on Google and ChatGPT. Gunning Fog Index (readability), a modified 5-question DISCERN Score (reliability), the Global Quality Score, and expert urologists' assessments guided responses' analysis. Independent t-tests let one evaluate Google's and ChatGPT's performance differences.

Results: ChatGPT showed notably better performance than Google in terms of the DISCERN Score (mean: 21.4 vs. 18.0; p = 0.002), Global Quality Score (mean: 4.8 vs. 4.0; p = 0.001), and expert urologists' ratings (mean: 4.6 vs. 3.8; p = 0.001). Though ChatGPT had a lower Gunning Fog Index (mean: 15.84 vs. 18.23), indicating more legible text, this difference was not statistically significant (p = 0.165).

Conclusion: ChatGPT exceeded Google in offering consistent, high-quality, expert-endorsed knowledge about kidney stones. These results show the possibilities of artificial intelligence language models in access to correct healthcare information and patient education. Additionally, improving transparency of information sources and implementing further verification mechanisms for clinical accuracy are recommended to ensure the reliability of AI-generated medical information.

Keywords: artificial intelligence, health literacy, internet, kidney stone, patient education, urolithiasis

Cite As: *Elmaagac B, Golbasi A, Ozercan AY, Bicer H. A Comparison Study of AI Response Quality in Kidney Stone-Related Questions: Google vs. ChatGPT. Endourol Bull. 2026;18(1):30-38.* https://doi.org/10.54233/endourolbull-1769794

Corresponding Author : *Burak Elmaağaç, MD. Department of Urology, Sağlık Bilimleri University, Kayseri City Hospital, Kayseri, Türkiye*
e-mail: burakelmaagac@gmail.com
Received : August 25, 2025 ***Accepted*** : November 8, 2025

ÖZET

Amaç: Türkiye'de böbrek taşıyla ilgili sık aranan sorularda Google ve ChatGPT'nin verdiği yanıtların okunabilirlik, güvenilirlik ve genel kalite açısından karşılaştırılması.

Gereç ve Yöntemler: Google Trends ile 2023 yılı için en sık aranan 10 soru ve ani yükseliş gösteren 4 konu belirlendi (toplam 14 sorgu). Google için ilk sayfa özetleri, ChatGPT (GPT-4) için doğrudan yanıtlar değerlendirildi. Okunabilirlik Gunning Fog İndeksi; güvenilirlik 5 maddelik modifiye DISCERN; genel kalite Global Quality Score ile puanlandı. Üç deneyimli üroloji uzmanı ayrıca klinik fayda ve doğruluk açısından 1–5 arası puan verdi.

Bulgular: ChatGPT, DISCERN (21,4 vs. 18,0; p = 0,002), Global Quality Score (4,8 vs. 4,0; p = 0,001) ve uzman değerlendirmelerinde (4,6 vs. 3,8; p = 0,001) Google'dan üstün bulundu. Okunabilirlik ChatGPT'de daha iyi olsa da (15,84 vs. 18,23) fark istatistiksel olarak anlamlı değildi (p = 0,165).

Sonuç: ChatGPT, böbrek taşıyla ilgili çevrim içi sağlık bilgisinin güvenilirliği ve genel kalitesi açısından Google'a göre daha tutarlı ve yüksek kaliteli içerik sunmaktadır. AI tabanlı dil modelleri hasta eğitimi ve sağlık okuryazarlığının desteklenmesinde potansiyele sahiptir; ancak kaynak şeffaflığının artırılması ve klinik doğruluk için ek doğrulama mekanizmaları önerilir.

Anahtar Kelimeler: böbrek taşları, hasta eğitimi, internet, sağlık okuryazarlığı, ürolitiyazis, yapay zekâ

INTRODUCTION

Patient education depends more and more on access to online health knowledge. For common yet complicated disorders like kidney stones, precise information is absolutely vital. People usually rely on search engines like Google nowadays to find health information. However, in this field, artificial intelligence (AI)-based language models, namely ChatGPT, have become a new substitute.

Google provides a vast pool of information, yet concerns exist regarding the accuracy, readability, and reliability of its content (1,2). AI-based language models such as ChatGPT increase accessibility and comprehension. ChatGPT produces coherent and contextually relevant replies, unlike conventional search engines that list several sources and demand users to filter and interpret information, hence perhaps lowering false information and improving patient understanding (3). In recent years, the impact of AI-based language models—especially for complex issues such as kidney diseases—has attracted significant attention. These models are noted not only for increasing the speed and scope of information access but also for enhancing patient education, providing clinical decision support, and improving the comprehensibility of medical knowledge. However, their occasional tendency to produce content that is inaccurate or not based on factual evidence—a phenomenon referred to as "hallucination"—highlights the need for cautious use of such systems. It has been shown that models like ChatGPT offer guideline-compliant and high-quality responses (4,5). Additionally, according to Google Trends data, internet searches related to kidney stones have been steadily increasing since 2004, reaching their peak in 2022 (6). Focusing on readability, reliability, general quality, and expert urologist reviews, this paper seeks to evaluate Google and ChatGPT in offering health information on kidney stones.

MATERIALS AND METHODS

Study Design and Data Source

This study evaluated the quality of the obtained information and assessed online search patterns of Turkish users connected to kidney stones using Google Trends. Conducted within Türkiye, the study examined Google search data ranging from January 1, 2023, to December 31, 2023.

Google Trends revealed four new searches with a dramatic rise in search frequency throughout the same period, and the top ten most often used search queries connected to "kidney stones". The investigation consisted of 14 search queries overall.

Top 10 Most Popular Search Queries

1. Kidney
2. Kidney stone
3. Kidney stone symptoms
4. Kidney pain
5. What causes kidney stones?
6. Kidney stone pain
7. Kidney stone surgery
8. How to pass a kidney stone?
9. What helps with kidney stones?
10. Kidney stone lithotripsy

Sudden Trending Queries During the Study Period

1. How to use horsetail herb for kidney stones?
2. What to eat to pass a kidney stone?
3. Does beer help pass kidney stones?
4. What relieves kidney stone pain?

Data Collection Process

The selected 14 queries were entered into both the Google search engine and the ChatGPT (GPT-4 model) platform. Google search results were compiled based on summaries of content appearing on the first page, while ChatGPT responses were directly retrieved from the platform. The collected data were then analyzed according to predefined evaluation criteria to assess readability, reliability, and quality of information.

Evaluation Criteria

The responses from Google and ChatGPT were evaluated based on four primary criteria:

1. Readability (Gunning Fog Index) (7)
2. The readability level of the responses was assessed using the Gunning Fog Index, which measures the years of formal education required to understand a given text. A lower index score indicates that the text is more accessible and easier to comprehend.
3. Reliability (Modified DISCERN Scale) (8)
4. The reliability of the information was assessed using a modified version of the DISCERN instrument, which consists of five key questions:
    - Is the study's objective clearly stated?
    - Are the information sources provided?
    - Is the presented information balanced and unbiased?
    - Does the content include up-to-date information?
    - Does the information assist patients in making informed decisions?
    - Each question was rated on a scale from 1 to 5, and a total reliability score was calculated.
5. Global Quality Score (9)
6. The overall quality of the responses was rated on a scale from 1 to 5 (1: Low quality, 5: High quality), considering readability, organization, and content richness.
7. Expert Urologist Review
8. The responses were evaluated by three experienced urology specialists (with a minimum of seven years of practice). The experts assessed the content based on scope, accuracy, and clinical usefulness, assigning a rating between 1 and 5 for each response.

Statistical Analysis

An independent t-test examined variations between Google's and ChatGPT's answers. A p-value of 0.05 was considered statistically significant. Every statistical analysis was done with Python version 3.9.

RESULTS

Gunning Fog Index

The Gunning Fog Index is a measure of text readability. GPT scored 15.84; the average Gunning Fog Index score for Google's content was 18.23. These numbers show that whilst GPT delivered knowledge more understandably, Google's material used a more sophisticated linguistic structure. The independent t-test results (t = -1.45, p = 0.165) revealed, nevertheless, that this variation was not statistically significant.

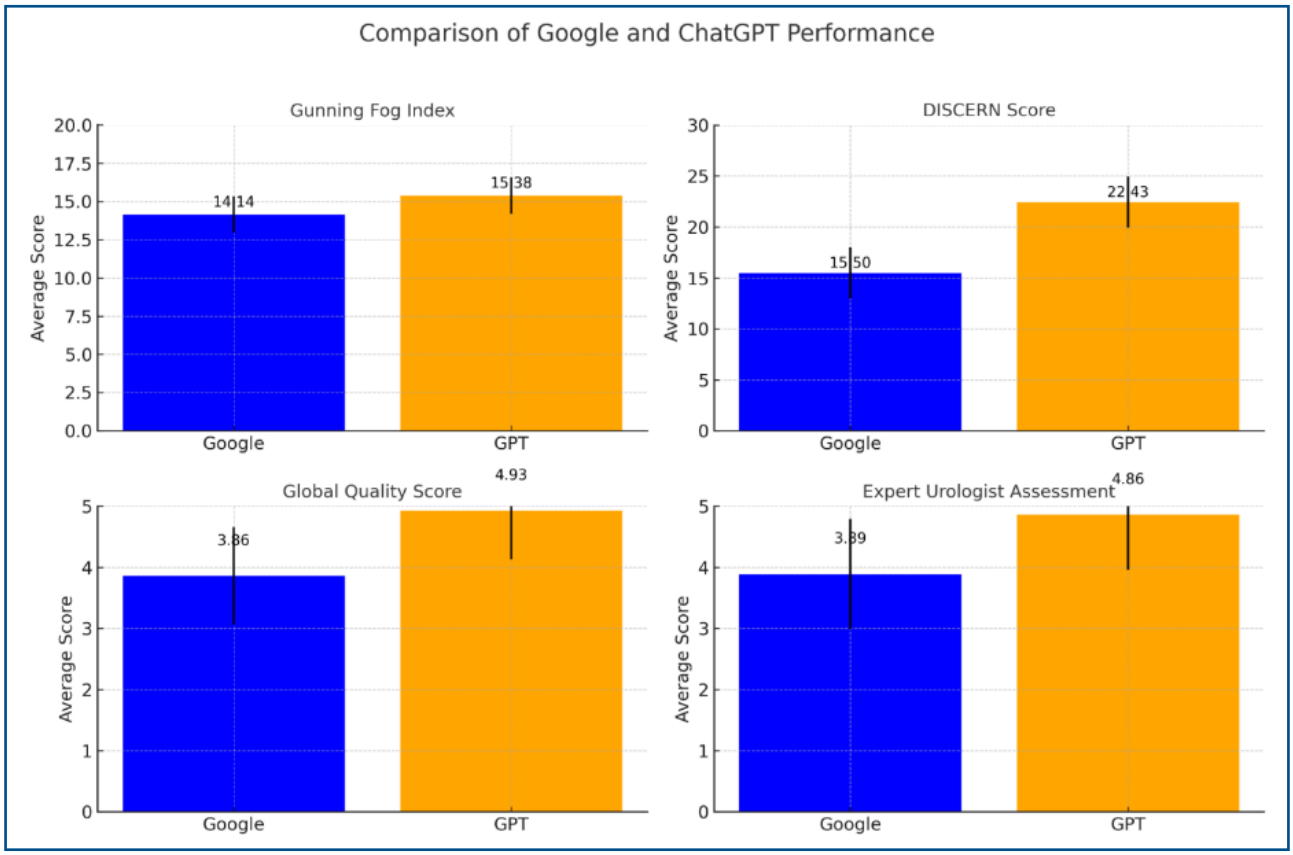

Figure 1. Comparison of readability (Gunning Fog Index), reliability (DISCERN Score), overall content quality (Global Quality Score), and expert urologist evaluation between Google and ChatGPT responses to kidney stone-related queries. ChatGPT consistently outperformed Google across all domains, with statistically significant differences for reliability, quality, and expert assessment ($p < 0.05$), whereas readability differences were not significant.

DISCERN Score

The DISCERN Score was used to assess the reliability of health information. Google's average DISCERN score was 18.0; GPT scored 21.4. With a statistically significant advantage in this criterion (t = -3.72, p = 0.002), GPT suggested offering more dependable and source-based health information.

Global Quality Score

The overall quality of the material was assessed using the Global Quality Score. GPT scored 4.8; the average score for Google material was 4.0. The t-test findings (t = -4.15, p = 0.001) showed that GPT beat Google noticeably in this regard, therefore confirming the idea that GPT-generated information is of better general quality.

Expert Urologist Review

Evaluations by seasoned urologists found that Google's material had an average score of 3.8, whereas GPT's material scored 4.6. With a t = -4.11, p = 0.001, this measure likewise revealed a statistically significant preference for GPT. Professional urologists thought GPT's comments were more accurate, relevant, and clinically helpful.

- **Gunning Fog Index Comparison:** This measure of text readability, Google and ChatGPT performed similarly, with ChatGPT showing a somewhat lower score, suggesting a rather more intelligible response structure.
- **DISCERN Score Comparison:** Assessing the reliability of information sources, ChatGPT demonstrated a significantly higher DISCERN score compared to Google, with a statistically significant difference.
- **Global Quality Score Comparison:** ChatGPT received a higher score than Google in terms of overall content quality, with this difference also being statistically significant.
- **Expert Urologist Assessment:** Based on evaluations by expert urologists, ChatGPT provided more accurate, reliable, and clinically relevant responses compared to Google, resulting in a higher overall score.

DISCUSSION

This study compares the responses provided by Google and ChatGPT to frequently asked kidney stone-related queries, evaluating readability, reliability, general quality, and expert assessment. The results indicate that ChatGPT outperformed Google in multiple key aspects, highlighting its potential as a valuable tool for disseminating health information. However, these findings should be considered in a broader context.

Google's Referral Policy and Approach to Questions

Google's investigation showed that search results regularly pointed people to Turkish private hospital websites. Particularly for the "kidney stone lithotripsy" search, consumers were frequently sent to promotional hospital pages instead of thorough, educational material. This could create commercial bias (10) and restrict access to trustworthy health information. By comparison, ChatGPT gave consumers a more all-encompassing viewpoint and a thorough description of lithotripsy and surgical techniques.

Response Length and Comprehensibility

Google's and ChatGPT's response lengths stood out as particularly different. Generally clear, direct, and straightforward Google answers let people get knowledge fast. Sometimes, though, this simplicity produced inadequate responses. Conversely, ChatGPT's answers were more thorough and instructive, which helped users to grasp kidney stones better. In our research, every search entered into ChatGPT produced answers giving consumers organized, all-encompassing knowledge. On the other hand, Google users sometimes found marketing material, which occasionally hampered their access to all the information.

Position in the Literature

Digital era access to health information has changed significantly. Previous research has thoroughly examined how well internet platforms provide health information, together with their drawbacks and obstacles. Even while Google is the most often used search engine for health-related searches, questions regarding information authenticity, reliability, and objectivity still exist. According to Eysenbach et al. (11), Google searches often direct consumers to commercial or partial content, therefore making dependable source access challenging. Morahan-Martin (2) further underlined that many sources of internet health information lack openness about their sources; therefore, users find difficulties assessing them.

Particularly, ChatGPT, an AI-based language model, constitutes a paradigm change in the distribution of health knowledge. Brown et al. (3) underlined that ChatGPT is a potential method for structured health information dissemination since it takes advantage of natural language processing capacity. According to Temsah et al. (12),

ChatGPT provides medical knowledge really well, but it needs further accuracy assurance through extra verification systems. Moreover, Nori et al. (13) evaluated AI models in the healthcare sector. They found that, although they improve patient-centered information accessibility, they also carry a risk of medical errors, which has to be reduced. In a recent study examining the compliance of AI-based applications with clinical guidelines, ChatGPT-4.0 and similar platforms demonstrated a high degree of adherence to the American Urological Association (AUA) guidelines in the management of vesicoureteral reflux (VUR) in children (14). This finding supports the potential use of AI as a clinical decision support tool. Additionally, large language models like ChatGPT offer healthcare professionals new opportunities in areas such as patient education, clinical decision support, and access to medical knowledge (15,16). Another important limitation of AI-based systems is the inherent risk of "hallucination," in which the model generates statements that appear plausible but are factually incorrect. This phenomenon underscores the need for ongoing validation, expert supervision, and external verification of AI-generated medical content to ensure clinical reliability. Supporting this, ChatGPT's responses to complex medical topics such as kidney transplantation were found to be largely accurate and sufficient based on expert evaluations (17,18).

Though ChatGPT has several benefits, our research found cases when the AI-generated answers included broad descriptions straying from the central question. Furthermore, direct contact with healthcare providers is still quite important for patients looking for medical guidance, considering Turkey's cultural focus on trust-based doctor-patient interactions.

Social and Commercial Influences on Search for Health Information

This study makes one of the major contributions by analyzing the social and commercial settings affecting popular health-related searches. For example, the search "How to use horsetail herb for kidney stones?" was often searched between June and July 2024, in line with a running advertisement campaign for this product in Turkey. This result implies that online search patterns capture not only actual health information requirements but also outside social and commercial factors. Future studies ought to investigate how online health information-seeking behavior is shaped by cultural and financial aspects.

Health Literacy and AI-Generated Content

The capacity of ChatGPT to provide thorough answers presents a benefit for encouraging health literacy. Topol (19) underlined how well artificial intelligence models reduce difficult medical subjects, therefore enabling the general public to find them more understandable. Nevertheless, too much detail could potentially overwhelm consumers. In this study, even if ChatGPT's thorough answers help to improve knowledge about kidney stones, some users could still want Google's simpler approach despite its restrictions. Recently, Bahçeci et al. conducted a comparative study evaluating the effectiveness of Microsoft Copilot and Google Search in answering patient inquiries about infertility (20). Their results showed that, similar to our findings, AI-based chatbots provided significantly higher understandability and actionability scores than Google Search responses, as assessed by the PEMAT-P tool. Although both platforms demonstrated readability levels above the recommended eighth-grade threshold, the AI model generated more comprehensive and patient-centered information. Together with the present research on urolithiasis, these complementary studies emphasize the growing role of large language models in improving online patient education and health literacy across different domains of urology.

Moreover, as artificial intelligence models continue to evolve rapidly, their performance and accuracy are expected to change over time. Future evaluations should therefore consider version-specific analyses to ensure the long-term applicability and reproducibility of AI-based findings in clinical communication and patient education.

Limitations

This study is among the first to evaluate Google and ChatGPT in the framework of Turkish health information searches, therefore providing an insightful analysis of digital health information availability. Still, certain restrictions have merit. The study focused on 14 frequently searched questions related to kidney stones, which may limit the statistical generalizability of the results. However, this focused and exploratory design ensured methodological consistency and enabled a detailed qualitative assessment by expert urologists, thereby offering a reliable baseline for future, larger-scale research. A key limitation is that the study relied only on search queries from the year 2023. Given the rapid evolution of Google and ChatGPT algorithms, the results may change over time, potentially affecting reproducibility and comparability in future analyses. Furthermore, constantly changing their algorithms and reactions, Google and ChatGPT could influence the repeatability of the findings over time. Additionally, during the conduct of this research, the release of ChatGPT-5 highlighted the rapid advancement of large language models. Such fast-paced developments may lead to notable differences in performance, content generation, and reliability across versions, potentially affecting the relevance of our current findings over time. Future research should therefore consider incorporating newer model versions and conducting longitudinal evaluations to capture the dynamic nature of these technologies and provide more up-to-date, generalizable insights.

CONCLUSION

According to this study, AI-based language models could be a great substitute for the distribution of health knowledge. Still, ChatGPT and related models ought to increase openness about their information sources. To improve the accessibility of accurate and objective health information, Google should thus hone its algorithm to lead people towards more balanced and neutral sources. Future studies should broaden this investigation by examining several health subjects using more extensive data sets.

Furthermore, more research might evaluate how diverse user profiles—especially those with varied degrees of education—interpret health information from several platforms. Such studies would offer a closer understanding of the efficiency of search engine-based and artificial intelligence-driven information.

Conflict of Interest: The authors declare that they have no conflict of interest.

Informed Consent: Not applicable (no human participants or identifiable data).

Funding: This research received no financial support.

Ethical Approval: This study is based solely on the analysis of publicly available online data; therefore, ethical committee approval was not required. As no human participants or personal data were involved, this study did not necessitate ethical approval according to institutional and journal policies.

Author Contributions:

- Conceptualization: Burak Elmaağaç, Abdullah Gölbaşı
- Methodology: Burak Elmaağaç, Ali Yasin Özercan
- Data Curation: Hüseyin Biçer, Ali Yasin Özercan
- Formal Analysis: Abdullah Gölbaşı
- Writing—Original Draft: Burak Elmaağaç
- Writing—Review & Editing: All authors
- Supervision: Burak Elmaağaç

REFERENCES

1. Eysenbach G. The impact of the Internet on health professionals and patients: pros and cons. *BMJ*. 2001;323(7327):731-3. https://doi.org/10.1136/bmj.323.7327.731
2. Morahan-Martin JM. How Internet users find, evaluate, and use online health information: a cross-cultural review. *Cyberpsychol Behav*. 2004;7(5):497-510. https://doi.org/10.1089/cpb.2004.7.497
3. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst*. 2020;33:1877-901.
4. Hanna DR, Creswell ML, Terry RS, Vergamini LB, Sardiu M, et al. Bing chat for kidney stone management questions based on the AUA guidelines: a comparison of chatbot conversation style modes. *World J Urol*. 2025;43(1):151. https://doi.org/10.1007/s00345-025-05533-4
5. Gençer Bingöl F, Ağagündüz D, Bingol MC. Accuracy of current large language models and the retrieval-augmented generation model in determining dietary principles in chronic kidney disease. *J Ren Nutr*. 2025;S1051-2276(25)00013-5. https://doi.org/10.1053/j.jrn.2025.01.004
6. Aiumtrakul N, Thongprayoon C, Suppadungsuk S, Krisanapan P, Pinthusopon P, et al. Global trends in kidney stone awareness: a time series analysis from 2004-2023. *Clin Pract (Basel)*. 2024;14(3):915-27. https://doi.org/10.3390/clinpract14030072
7. Gunning R. *The technique of clear writing*. New York: McGraw-Hill; 1952.
8. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. *J Epidemiol Community Health*. 1999;53(2):105-11. https://doi.org/10.1136/jech.53.2.105
9. Bernard A, Langille M, Hughes S, Rose C, Leddin D, Veldhuyzen van Zanten S. *A systematic review of patient inflammatory bowel disease information on the Internet.* Am J Gastroenterol. 2007;102(9):2070–77. https://doi.org/10.1111/j.1572-0241.2007.01325.x
10. Eysenbach G, Köhler C. How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. *BMJ*. 2002;324(7337):573-7. https://doi.org/10.1136/bmj.324.7337.573
11. Eysenbach G. Consumer health informatics. *BMJ*. 2000;320(7251):1713-6. https://doi.org/10.1136/bmj.320.7251.1713
12. Temsah MH, Aljamaan F, Malki KH, Alhasan K, Altamimi I, et al. ChatGPT and the future of digital health: a study on healthcare workers' perceptions and expectations. *Healthcare (Basel)*. 2023;11(13):1812. https://doi.org/10.3390/healthcare11131812
13. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. *arXiv Preprint*. 2023. https://doi.org/10.48550/arXiv.2303.13375
14. Sarikaya M, Ozcan Siki F, Ciftci I. Use of artificial intelligence in vesicoureteral reflux disease: a comparative study of guideline compliance. *J Clin Med*. 2025;14(7):2378. https://doi.org/10.3390/jcm14072378
15. Jo E, Song S, Kim JH, Lim S, Kim JH, et al. Assessing GPT-4's performance in delivering medical advice: comparative analysis with human experts. *JMIR Med Educ*. 2024;10:e51282. https://doi.org/10.2196/51282
16. Vrdoljak J, Boban Z, Vilović M, Kumrić M, Božić J. A review of large language models in medical education, clinical decision support, and healthcare administration. *Healthcare (Basel)*. 2025;13(6):603. https://doi.org/10.3390/healthcare13060603
17. Çolakoğlu Y, Ayten A, Sertkaya C, Toksal K, Karadağ S. Evaluation of Chat generative pretrained transformer (ChatGPT) performance in answering kidney transplant related questions. *New J Urol*. 2025;20(1):21-3. https://doi.

org/10.33719/nju1613084

18. Lee J, Park J, Han HS. Using ChatGPT for kidney transplantation: perceived information quality by race and education levels. *Clin Transplant*. 2024;38(7):e15378. https://doi.org/10.1111/ctr.15378
19. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44-56. https://doi.org/10.1038/s41591-018-0300-7
20. Bahçeci T, Elmaağaç B, Ceyhan E. Comparative analysis of the effectiveness of Microsoft Copilot artificial intelligence chatbot and Google Search in answering patient inquiries about infertility: evaluating readability, understandability, and actionability. IJIR: Your Sexual Medicine Journal. 2025. https://doi.org/10.1038/s41443-025-01056-z