Available online at www.medicinescience.org

ORIGINAL ARTICLE

Medicine Science International Medical Journal

Medicine Science 2026;15(1):243-8

Assessment of ChatGPT-4o's answers to common questions on thyroid fine-needle aspiration biopsy

Ali Salbas[1], Rasit Eren Buyuktoka[2]

[1]*İzmir Katip Celebi University, Ataturk Training and Research Hospital, Department of Radiology, İzmir, Türkiye*
[2]*Foca State Hospital, Department of Radiology, İzmir, Türkiye*

Received 20 September 2025; Accepted 16 October 2025
Available online 23 February 2026 with doi: 10.5455/medscience.2025.09.250

Abstract

This study set out to assess the quality of ChatGPT-4o's (Chat Generative Pre-trained Transformer, version 4o) replies to common patient questions concerning thyroid fine-needle aspiration biopsy (FNAB). A cross-sectional design was employed, in which patient-focused questions were gathered using the search phrase "frequently asked questions about thyroid biopsy" on Google. Following the removal of duplicates and overlapping items, 20 unique questions were chosen. Each question was submitted to ChatGPT-4o in a new session. The generated responses were then evaluated by 12 radiologists, all blinded to the source of the answers. Ratings were given on a 5-point Likert scale across four categories: relevance, accuracy, clarity, and completeness. Descriptive analyses were performed, and interrater reliability was calculated using the intraclass correlation coefficient (ICC). All 20 questions received scores between 3 and 5 in every category. The overall mean score was 4.72±0.12. Relevance achieved the best performance with a mean of 4.95±0.06, while clarity was the lowest at 4.61±0.23. The reliability analysis showed weak agreement among evaluators, with ICC values of –0.028 (p=0.863) for relevance, 0.061 (p=0.005) for accuracy, 0.072 (p=0.002) for clarity, 0.031 (p=0.016) for completeness, and 0.061 (p=0.002) for the overall score. In conclusion, ChatGPT-4o produced highly relevant, accurate, and generally comprehensive responses to patient inquiries regarding thyroid FNAB. Nonetheless, the limited interrater reliability underscores variability in expert judgment, especially in clarity and completeness. Although ChatGPT-4o holds promise as a supportive tool for patient education, its outputs should be reviewed and tailored by healthcare professionals prior to use in clinical practice.

Keywords: Thyroid nodule, biopsy, fine-needle, patient education, natural language processing, artificial intelligence

Introduction

Large language models (LLMs) are advanced artificial neural network systems trained on vast amounts of textual data to generate human-like language outputs. LLM-based chatbots such as ChatGPT (Chat Generative Pre-trained Transformer) have started to attract attention in the medical field due to their text generation capabilities [1]. In particular, these models have been investigated in radiology for diagnostic support and for optimizing tasks such as report generation and formatting [2–5]. Beyond these applications, LLMs have also been explored for enhancing clinical communication and supporting radiology education, thereby broadening their potential role in medical practice [6,7].

Biopsy is a commonly used method in the diagnostic process and plays an important role in the diagnosis of various diseases [8–10]. Thyroid fine-needle aspiration biopsy (FNAB) is a widely used, minimally invasive method for evaluating thyroid nodules in terms of malignancy [9,11]. However, despite its minimally invasive nature, this procedure may cause anxiety and uncertainty in many patients. Today, patients largely seek health-related information online, and artificial intelligence (AI)-based chatbots are becoming increasingly prominent in this context. The potential of models such as ChatGPT to provide reliable information in patient education has been extensively investigated in both radiology and other fields of medicine [12–15].

To the best of our knowledge, there is no study in the English literature that evaluates the quality of responses provided by large language models to frequently asked patient questions specifically about thyroid biopsy. In this context, our study is the first to evaluate ChatGPT-4o's responses to frequently asked

CITATION
Salbas A, Buyuktoka RE. Assessment of ChatGPT-4o's answers to common questions on thyroid fine-needle aspiration biopsy. Med Science. 2026;15(1):243-8.

Corresponding Author: Ali Salbas, İzmir Katip Celebi University, Ataturk Training and Research Hospital, Department of Radiology, İzmir, Türkiye
Email: dralisalbas@gmail.com

DOI: 10.5455/medscience.2025.09.250

patient questions about thyroid FNAB. In this regard, our study aims to highlight the potential of large language models in patient education processes.

Material and Methods

Ethical approval was obtained from the Health Research Ethics Committee of Izmir Katip Celebi University (Date: September 11, 2025; Decision No: 0488). As no patient data or human participants were involved, informed consent was not required. This cross-sectional study was conducted to evaluate the quality of ChatGPT-4o's responses to commonly asked patient questions regarding thyroid FNAB. To identify these questions, a new Google (Alphabet Inc., Mountain View, CA, USA) account with no prior search history was created. By searching the phrase 'frequently asked questions about thyroid biopsy,' a total of 200 initial questions were identified from the 'Other questions' section on the main screen on September 12, 2025 [16]. Two board-certified radiologists with experience in thyroid imaging and biopsy reviewed the questions and reached a consensus to eliminate duplicate or semantically similar items. As a result, 20 unique patient-centered questions were selected for analysis (Table 1). Questions were included as originally phrased in the Google search results to reflect real-world patient language.

Table 1. Patient-centered questions selected for ChatGPT-4o evaluation regarding thyroid fine-needle aspiration biopsy

No	Question
1	What are the do's and don'ts after a thyroid biopsy?
2	How serious is a thyroid biopsy?
3	How long does it take to heal from a thyroid biopsy?
4	Can I drive myself after a thyroid biopsy?
5	How painful is a thyroid biopsy?
6	What problems can occur after a thyroid biopsy?
7	Can thyroid biopsies be repeated?
8	Can I drink water before a thyroid biopsy?
9	Under what circumstances is a thyroid biopsy performed?
10	Can performing a thyroid biopsy cause the cancer to spread?
11	Is swelling normal after a thyroid biopsy?
12	What are the signs of infection after a thyroid biopsy?
13	Can I eat after a thyroid biopsy?
14	What type of doctor does thyroid nodule biopsy?
15	Is the needle used in a thyroid biopsy thick?
16	Can I take a shower after a thyroid biopsy?
17	Is a biopsy performed on every thyroid nodule?
18	Is any special preparation required before a thyroid biopsy?
19	Are stitches required after a thyroid biopsy?
20	Do they numb you before a thyroid biopsy?

The selected questions were individually submitted to ChatGPT-4o (OpenAI Inc., San Francisco, CA, USA) using an account that had not been used for any prior interactions. The model was accessed via its official web interface [17]. No additional prompts or contextual information were included. Only the questions were entered. To prevent the model from modifying its response behavior based on prior prompts within the same session, a phenomenon referred to as in-context adaptation, a new session was initiated for each question by clearing the chat history [18,19]. All questions were submitted on September 12–14, 2025.

The responses generated by ChatGPT-4o were then evaluated by 12 radiologists with 5 to 13 years of experience in thyroid FNAB. The evaluators were blinded to the origin of the responses and were informed only that they were assessing answers to patient questions about thyroid biopsy. Each response was rated across four criteria: relevance, accuracy, clarity, and completeness. Each response was scored from 1 to 5 for each criterion, with higher scores indicating better performance [20].

The rating scale was adapted from two previously published studies [14,20]. Each evaluation criterion was defined as follows:

- **Relevance:** Whether the answer directly addressed the question posed.
- **Accuracy:** Whether the information provided was factually correct and medically appropriate.
- **Clarity:** Whether the response was well-structured, easy to understand, and free from ambiguity.
- **Completeness:** Whether the answer included all necessary information to adequately address the question.

Statistical Analysis

All data were analyzed using IBM SPSS Statistics version 26.0 (IBM Corp., Armonk, NY, USA). The results were reported as means and standard deviations. Interrater reliability (IRR) was assessed using the intraclass correlation coefficient (ICC) for each of the four evaluation criteria. A two-way random effects model with absolute agreement was applied. Ninety-five percent confidence intervals (95% CI) and p-values were calculated, and a p-value <0.05 was considered statistically significant.

Result

A total of 20 questions were evaluated by 12 radiologists across four criteria: relevance, accuracy, clarity, and completeness. Each criterion was scored on a 5-point scale. The lowest score assigned by the evaluators for any response was 3 out of 5 in each subcategory of the rating scale. Among the four criteria, relevance received the highest mean score, while clarity showed the lowest (Table 2). The overall mean score was 4.72±0.12. The mean scores and corresponding standard deviations for each criterion are illustrated in Figure 1. An example of a patient question and ChatGPT-4o's response is presented in Figure 2.

DOI: 10.5455/medscience.2025.09.250

Table 2. Descriptive statistics of scores

Evaluation criteria	Number of Questions	Mean±SD
Relevance	20	4.95±0.06
Accuracy	20	4.65±0.19
Clarity	20	4.61±0.23
Completeness	20	4.69±0.13
Total	80	4.72±0.12

SD: Standard deviation

Table 3. Analysis of inter-rater reliability

Evaluation criteria	ICC	Lower bound (95% CI)	Upper bound (95% CI)	p
Relevance	–0.028	–0.159	0.104	0.863
Accuracy	0.061	–0.071	0.191	0.005
Clarity	0.072	–0.060	0.201	0.002
Completeness	0.031	–0.101	0.161	0.016
Total	0.061	–0.071	0.190	0.002

CI: Confidence interval; ICC: Intraclass correlation coefficient

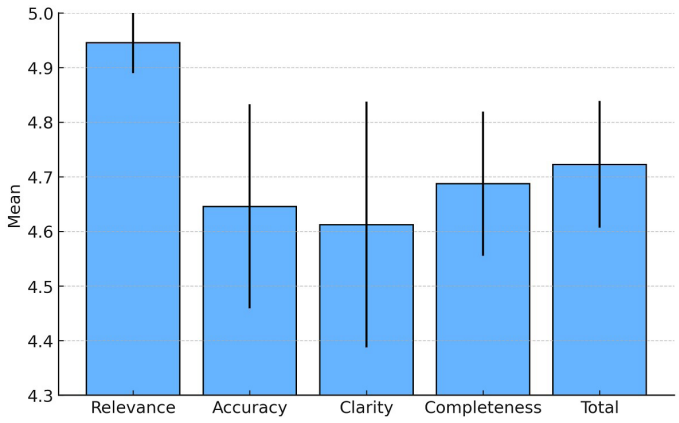

Figure 1. Graph of average scores and standard deviations

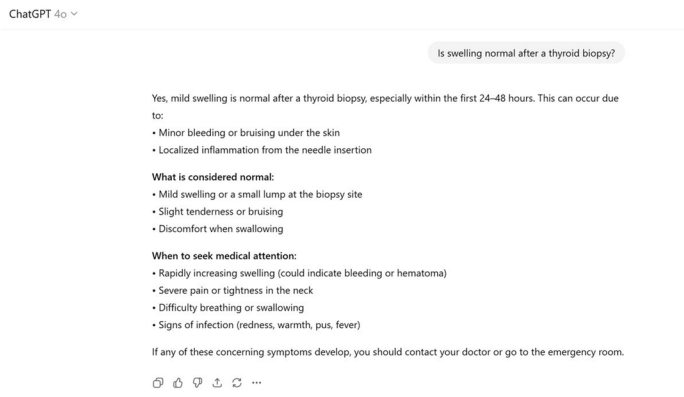

Figure 2. Example of a patient question and ChatGPT-4o's corresponding response regarding thyroid fine-needle aspiration biopsy

The interrater reliability analysis demonstrated low agreement among the evaluators. The ICC values were –0.028 for relevance, 0.061 for accuracy, 0.072 for clarity, 0.031 for completeness, and 0.061 for the total scores (Table 3) (Figure 3). While clarity showed the highest ICC, relevance demonstrated the lowest. Statistically significant results were observed for accuracy (P=.005), clarity (P=.002), completeness (P=.016), and total score (P=.002), whereas relevance did not reach significance (P=.863). Despite the statistical significance for accuracy, clarity, completeness, and total, the overall ICC values indicated poor interrater reliability.

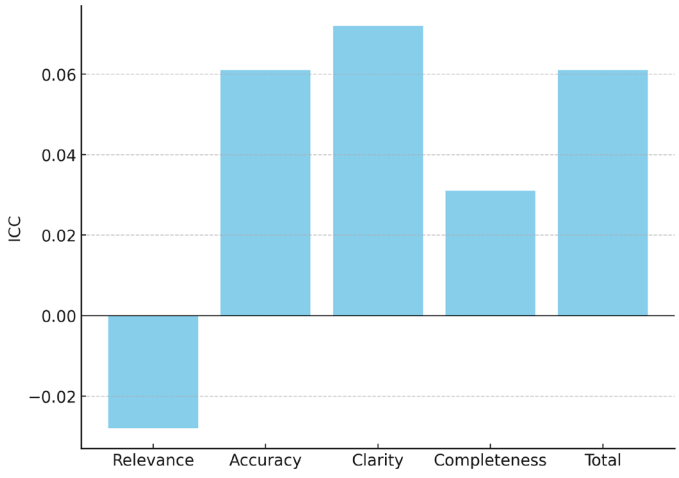

Figure 3. Bar chart of intraclass correlation coefficients (ICC) for relevance, accuracy, clarity, completeness, and total scores

DOI: 10.5455/medscience.2025.09.250

Discussion

In this study, we evaluated ChatGPT-4o's responses to frequently asked patient questions regarding thyroid fine needle aspiration biopsy using four predefined criteria: relevance, accuracy, clarity, and completeness. Our findings demonstrated that the overall quality of responses was high, with a mean score of 4.72±0.12. Among the four criteria, relevance received the highest mean score, while clarity showed the lowest. Although the mean scores indicated generally favorable performance, the interrater reliability analysis revealed low agreement among the evaluators.

Currently, patients can access information about medical procedures and diseases through the internet and artificial intelligence–based language models. While this situation has positive aspects, it has also been reported that the acquired information may increase patient anxiety [21]. In a recent survey conducted in the United States, 53% of participants stated that they would trust artificial intelligence in treatment planning [22].

FNAB is a widely used minimally invasive procedure in the management of thyroid nodules. However, despite being minimally invasive, FNAB may still cause pain and anxiety in patients [23,24]. Similarly, it has been reported that patients scheduled for biopsy procedures experience increased emotional stress and elevated serum cortisol levels prior to the intervention [25]. It has also been demonstrated that patients' pre- and post-procedural anxiety can be reduced through appropriate educational methods. For example, in a study on thyroid FNAB, patients who received video-based education had lower anxiety scores compared with those who were provided with written information alone [26]. However, to date, no directly comparable study has been identified in the literature that evaluates the responses of large language models such as ChatGPT-4o to frequently asked patient questions regarding FNAB. In contrast, studies conducted in different clinical domains have examined the performance of LLMs in responding to patient questions from various perspectives. Gordon et al. reported that in their study evaluating ChatGPT's responses to patient questions regarding radiological imaging, which were assessed by both radiologists and patient advocates, 83% of the responses were accurate, while 99% were at least partially relevant [7]. Another study demonstrated that ChatGPT-3.5 provided correct answers to 70.8% of questions related to lung cancer screening, although it also produced errors in certain cases [27]. In a study evaluating the use of ChatGPT for preparing patient brochures on interventional radiology procedures, the model was able to convey basic information but also exhibited notable omissions and errors in certain cases. Interestingly, while no major inaccuracies were identified for procedures such as thyroid FNAB, multiple errors were reported in brochures related to lung ablation and lung biopsy [28]. In a study focusing on thyroid nodules, approximately 69% of the responses were reported to be accurate or partially accurate, whereas another study evaluating patient questions about thyroid cancer found that only 52% of the responses were considered satisfactory [13,29]. In two studies conducted on thyroid surgery and thyroid nodules, ChatGPT was reported to generally provide accurate information, although its language was often complex and not patient-friendly [30,31].

In our study, ChatGPT-4o's responses to frequently asked questions regarding FNAB were generally rated highly, with a mean score of 4.72 out of 5. This finding differs from some previous research, which reported lower levels of accuracy and perceived adequacy. These differences may largely be related to the nature of the clinical topics evaluated and the formulation of the questions. In areas such as FNAB, which require more limited and standardized information, the model's performance appears to be higher, whereas in more complex fields such as oncology or surgery, ensuring both accuracy and patient-friendly language can be more challenging. In addition, the developmental level of the versions used and the evaluation criteria applied may vary across studies, which can influence the reported outcomes.

Although the responses of ChatGPT-4o in our study generally received high scores, the low level of agreement among evaluators (as reflected by low ICC values) was noteworthy. This finding suggests that subjective differences in interpretation may have influenced the evaluation of ChatGPT's responses. While all evaluators scored the answers using the same criteria, factors such as the level of detail or style of expression may have been interpreted differently. Similar studies in the literature have also reported low agreement among evaluators, suggesting that the assessment of LLM-generated responses using standardized metrics may present certain challenges [14,20]. The fact that medical experts, even when given clear instructions, often disagree when rating the quality of a language model's responses should not necessarily be interpreted as a shortcoming of the evaluators. Rather, it reflects the complexity of medicine, where there may be multiple valid ways to explain a procedure and where different experts may emphasize different aspects.

A study examining frequently asked questions about radiotherapy also reported low interrater agreement. In this study, the responses of GPT-4 and GPT-3.5 were found to be largely similar and consistent in terms of content; however, readability analyses indicated that both versions required a university-level literacy [32]. Similarly, in our study, ChatGPT-4o's clarity score was the lowest among the evaluated criteria (4.61±0.23). The relatively lower clarity scores may reflect the inherent difficulty of translating complex medical terminology into simpler, patient-friendly expressions. This finding indicates that large language models should not only be capable of generating accurate and comprehensive responses but also be developed to use patient-friendly and easily understandable language.

The fact that the ChatGPT-4o responses evaluated in our study received relatively high average scores and that none of the answers were classified as inadequate demonstrates that the model was generally able to provide accurate, relevant, and explanatory responses to basic patient questions regarding

DOI: 10.5455/medscience.2025.09.250

thyroid FNAB. However, as emphasized in the literature, one of the major limitations of large language models is the risk of producing information that is presented in a convincing manner but is, in fact, incorrect, a phenomenon known as hallucination [33–35]. In addition, issues such as fabricating references or overlooking critical medical details also raise serious concerns regarding patient safety [33]. Therefore, for LLMs to be used optimally in practice, it is essential that their responses are reviewed and verified by healthcare professionals and, when necessary, adapted to the patient's individual circumstances. This approach may not only help alleviate patients' procedure-related anxiety but also enhance the reliability of the information provided, thereby supporting the patient education process more effectively. In clinical practice, ChatGPT-4o may also assist physicians in patient communication and the preparation of educational materials for FNAB, potentially helping to reduce pre-procedural anxiety.

This study has several limitations. First, the evaluations were conducted by 12 radiologists, which may limit generalizability. Including experts from other specialties and patient advocates could provide broader perspectives, while future studies incorporating patient input and readability analyses (e.g., Flesch–Kincaid) may better assess clarity. Although mean scores were high, interrater reliability was low, likely reflecting minor differences in expert interpretation rather than deficiencies in response quality. Second, the questions were derived solely from Google searches, which may not represent the full range of patient inquiries across regions or platforms such as forums and social media. Third, the analysis was limited to ChatGPT-4o; other large language models might yield different results, and the findings reflect this model's performance at a specific point in time. Finally, only English-language questions and four evaluation criteria were analyzed, excluding other potentially relevant aspects such as consistency or evidence-based content. Despite these limitations, to the best of our knowledge, this is the first study to evaluate ChatGPT-4o's responses to frequently asked patient questions regarding thyroid fine needle aspiration biopsy. Our findings contribute to the growing body of literature on the role of large language models in patient education.

Future studies should aim to overcome these limitations by including a larger and more diverse group of evaluators from different specialties and by incorporating the perspectives of patients themselves. In addition, expanding the pool of questions to include content from multiple platforms and different languages could provide a more comprehensive assessment of patient concerns. Finally, prospective studies investigating the real-world impact of AI-generated responses on patient understanding, satisfaction, and clinical outcomes are warranted.

Conclusion

This study evaluated ChatGPT-4o's responses to frequently asked patient questions about thyroid fine needle aspiration biopsy using four predefined criteria. Although interrater reliability was low, reflecting variability in expert assessments, ChatGPT-4o demonstrated considerable potential as a supplementary tool in patient education. With further refinement and cautious integration into clinical practice, large language models may serve as valuable adjuncts to improve patient understanding of radiological procedures. Nonetheless, information provided by large language models should not replace professional medical advice.

Acknowledgments
We sincerely thank all participants who contributed to the conduct of this study.

Conflict of Interests
The authors declare that there is no conflict of interest in the study.

Financial Disclosure
The authors declare that they have received no financial support for the study.

Ethical Approval
The study received ethical approval from the Health Research Ethics Committee of İzmir Katip Celebi University (Date: September 11, 2025; Decision No: 0488).

References

1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. Nat Med. 2023;29:1930-0.
2. Ozenbas C, Engin D, Altinok T, et al. ChatGPT-4o's performance in brain tumor diagnosis and MRI findings: A comparative analysis with radiologists. Acad Radiol. 2025;32:3608-17.
3. Fink MA, Bischoff A, Fink CA, et al. Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer. Radiology. 2023;308:e231362.
4. Mitsuyama Y, Tatekawa H, Takita H, et al. Comparative analysis of GPT-4-based ChatGPT's diagnostic performance with radiologists using real-world radiology reports of brain tumors. Eur Radiol. 2024;35:1938-47.
5. Dehdab R, Brendlin A, Werner S, et al. Evaluating ChatGPT-4V in chest CT diagnostics: a critical image interpretation assessment. Jpn J Radiol. 2024;42:1168-77.
6. Lyo SK, Cook TS. Integrating large language models into radiology education: an interpretation-centric framework for enhanced learning while supporting workflow. J Am Coll Radiol. 2025;22:1271-80.
7. Gordon EB, Towbin AJ, Wingrove P, et al. Enhancing patient communication with Chat-GPT in radiology: evaluating the efficacy and readability of answers to common imaging-related questions. J Am Coll Radiol. 2024;21:353–59.
8. Kubo T, Furuta T, Johan MP, et al. A meta-analysis supports core needle biopsy by radiologists for better histological diagnosis in soft tissue and bone sarcomas. Medicine (Baltimore). 2018;97:e11567.
9. Koc AM, Adibelli ZH, Erkul Z, et al. Comparison of diagnostic accuracy of ACR-TIRADS, American Thyroid Association (ATA), and EU-TIRADS guidelines in detecting thyroid malignancy. Eur J Radiol. 2020;133:109390.
10. Yasar Y, Coskun M, Yasar E, et al. Optimization of pediatric kidney biopsy: Impact of needle size and passes. Acad Radiol. 2025;32:3659-66.
11. Koc AM, Adibelli ZH, Erkul Z, Sahin Y. Evaluation of fine needle aspiration biopsy (FNAB) results in macrocalcified thyroid nodules. J Tepecik Educ Res Hosp. 2021;31:103-9.
12. Scheschenja M, Viniol S, Bastian MB, et al. Feasibility of GPT-3 and GPT-4 for in-depth patient education prior to interventional radiological procedures: A comparative analysis. Cardiovasc Intervent Radiol. 2024;47:245-50.
13. Campbell DJ, Estephan LE, Sina EM, et al. Evaluating ChatGPT responses

DOI: 10.5455/medscience.2025.09.250

on thyroid nodules for patient education. Thyroid. 2024;34:371-7.

14. Magruder ML, Rodriguez AN, Wong JCJ, et al. Assessing ability for ChatGPT to answer total knee arthroplasty-related questions. J Arthroplasty. 2024;39:2022-7.

15. Erkaya R, Atalar S. Evaluating the accuracy and reliability of ChatGPT in breastfeeding counseling: A literature-based assessment. Med-Science. 2025;14:745.

16. Google. Google. http://www.google.com access date 24.09.2025

17. OpenAI. ChatGPT-4o web platform. http://chat.openai.com access date 05.07.2025

18. Gotta J, Le Hong QA, Koch V, et al. Large language models (LLMs) in radiology exams for medical students: Performance and consequences. Rofo. 2025;197:1057-67.

19. OpenAI. GPT-4o System Card. https://openai.com/index/gpt-4o-system-card/ access date 01.08.2025

20. Ayık G, Ercan N, Demirtaş Y, et al. Evaluation of ChatGPT-4o's answers to questions about hip arthroscopy from the patient perspective. Jt Dis Relat Surg. 2025;36:193-9.

21. Huisman M, Joye S, Biltereyst D. Searching for health: doctor Google and the shifting dynamics of the middle-aged and older adult patient–physician relationship and interaction. J Aging Health. 2020;32:998-1007.

22. SWNS. Seventy-five percent of Americans think they understand their health better than doctors. New York Post. https://nypost.com/2023/12/06/lifestyle/75-of-americans-think-they-understand-their-health-better-than-doctors/ access date 21.08.2025

23. Barlas T, Sodan HN, Avci S, et al. The impact of classical music on anxiety and pain perception during a thyroid fine needle aspiration biopsy. Hormones (Athens). 2023;22:581-5.

24. Gürkan O, Kaya MF. Effect of music on anxiety and pain levels of patients undergoing thyroid fine needle aspiration biopsy: A randomized controlled study. Acad Radiol. 2024;31:538-43.

25. Gustafsson O, Theorell T, Norming U, et al. Psychological reactions in men screened for prostate cancer. Br J Urol. 1995;75:631-6.

26. Ay S, Ata N, Oncu F. Effect of an information video before thyroid biopsy on patients' anxiety. J Invest Surg. 2022;35:531-4.

27. Rahsepar AA, Tavakoli N, Kim GHJ, et al. How AI responds to common lung cancer questions: ChatGPT versus Google Bard. Radiology. 2023;307:e230922.

28. Kooraki S, Hosseiny M, Jalili MH, et al. Evaluation of ChatGPT-generated educational patient pamphlets for common interventional radiology procedures. Acad Radiol. 2024;31(11):4548-4553. doi:10.1016/j.acra.2024.05.024

29. Gorris MA, Randle RW, Obermiller CS, et al. Assessing ChatGPT's capability in addressing thyroid cancer patient queries: a comprehensive mixed-methods evaluation. J Endocr Soc. 2025;9:bvaf003.

30. Köroğlu EY, Fakı S, Beştepe N, et al. A novel approach: evaluating ChatGPT's utility for the management of thyroid nodules. Cureus. 2023;15:e47576.

31. Sahin S, Tekin MS, Yigit YE, et al. Evaluating the success of ChatGPT in addressing patient questions concerning thyroid surgery. J Craniofac Surg. 2024;35:e572-e575.

32. Grilo A, Marques C, Corte-Real M, et al. Assessing the quality and reliability of ChatGPT's responses to radiotherapy-related patient queries: Comparative study with GPT-3.5 and GPT-4. JMIR Cancer. 2025;11:e63677.

33. Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting hallucinations in large language models using semantic entropy. Nature. 2024;630:625-30.

34. Nakaura T, Ito R, Ueda D, et al. The impact of large language models on radiology: A guide for radiologists on the latest innovations in AI. Jpn J Radiol. 2024;42:685-96.

35. Salbas A, Buyuktoka RE. Performance of large language models in recognizing brain MRI sequences: A comparative analysis of ChatGPT-4o, Claude 4 Opus, and Gemini 2.5 Pro. Diagnostics (Basel). 2025;15:1919.