

Evaluating the Success Rate of the Online Chat-based Artificial Intelligence Program ChatGPT in Answering Basic Questions Related to Thyroid Cancer

Çevrimiçi Sohbet Tabanlı Yapay Zeka Programı ChatGPT'nin Tiroid Kanseri ile İlgili Temel Soruları Yanıtlamadaki Başarı Oranının Değerlendirilmesi

Yiğit Türk¹, Bahadır Emre Baki¹, Ahmet Cem Dural², Serkan Teksöz³, Özer Makay^{4,5}, Recep Gökhan İçöz¹, Murat Özdemir¹

¹Ege University Faculty of Medicine, Department of General Surgery, İzmir, Türkiye

²İstinye University Faculty of Medicine, Department of General Surgery, İstanbul, Türkiye

³İstanbul University-Cerrahpaşa, Cerrahpaşa Faculty of Medicine, Department of General Surgery, İstanbul, Türkiye

⁴Özel Sağlık Hospital, Centre for Endocrine Surgery, İzmir, Türkiye

⁵Aristotle University Faculty of Medicine, Department of General Surgery, Thessaloniki, Greece

Cite as: Türk Y, Baki BE, Dural AC, et al. Evaluating the success rate of the online chat-based artificial intelligence program ChatGPT in answering basic questions related to thyroid cancer. Anatol J Gen Med Res. 2025;35(1):83-9

Abstract

Objective: ChatGPT, an advanced conversational bot based on artificial intelligence (AI) and a large language model, is designed to understand and generate responses to inputs. This study aims to assess the accuracy of responses provided by ChatGPT to questions that might be asked by patients concerning thyroid cancer.

Methods: A total of 27 questions in Turkish, relevant to thyroid cancer and likely to be asked by non-healthcare professionals, were prepared under four categories (general information, diagnosis, treatment, follow-up). These questions were posed to the free public version of ChatGPT, version 3.5. Three experts in endocrine surgery (A.C.D., S.T., Ö.M.) were asked to evaluate the responses. The answers were classified into three categories: appropriate, inappropriate, and insufficient/incomplete.

Results: Upon evaluating the responses given by ChatGPT to the prepared questions across the four categories, 9 responses (33.3%) were considered "appropriate" by two of the three experts and "insufficient/incomplete" by one expert. Six responses (22.2%) were deemed "appropriate" by two experts and "inappropriate" by one. Overall, a total of 16 responses (59.25%) were considered "appropriate" by at least two experts.

Conclusion: At this stage, AI-based conversational programs like ChatGPT are not seen as capable of replacing a specialist from whom patients receive medical advice.

Keywords: ChatGPT, thyroid cancer, artificial intelligence



Address for Correspondence/Yazışma Adresi: Yiğit Türk MD, Ege University Faculty of Medicine, Department of General Surgery, İzmir, Türkiye
E-mail: yigit.turk@ege.edu.tr
ORCID ID: orcid.org/0000-0001-9755-8163

*Our paper was presented as an oral presentation at the 23rd National Surgery Congress held in Antalya between April 24-28.

Received/Geliş tarihi: 17.08.2024

Accepted/Kabul tarihi: 25.03.2025

Epub: 29.04.2025

Published date/Yayınlanma tarihi: 15.05.2025



Copyright© 2025 The Author. Published by Galenos Publishing House on behalf of University of Health Sciences Turkey, İzmir Tepecik Education and Research Hospital. This is an open access article under the Creative Commons AttributionNonCommercial 4.0 International (CC BY-NC 4.0) License.

Öz

Amaç: Yapay zeka (YZ) tabanlı bir büyük dil modeli olan ChatGPT, girdileri anlayabilen ve üretebilen bir program olan gelişmiş bir sohbet botudur. Bu çalışmada tiroid kanseri ile ilgili hastaların sorabileceği sorulara ChatGPT'nin verdiği yanıtların doğruluğunu değerlendirmeyi amaçladık.

Yöntem: Tiroid kanseri ile ilgili sağlık profesyoneli olmayanların sorabileceği toplam dört başlıkta (genel bilgiler, tanı, tedavi, takip) 27 Türkçe soru hazırlandı. Bu sorular ücretsiz halka açık versiyon olan ChatGPT sürüm 3.5'e soruldu. Endokrin cerrahisi alanında uzman üç akademisyenden (A.C.D., S.T., Ö.M.) yanıtları değerlendirmesi istendi. Cevaplar uygun/uygun değil/ yeterli değil-eksik, şeklinde üç ayrı grupta sınıflandırdı.

Bulgular: ChatGPT'nin hazırlanan dört gruptaki sorulara verdiği yanıtlar genel olarak değerlendirildiğinde 27 soruya verilen 9 yanıt (%33,3), üç uzmandan ikisi tarafından "uygun", bir uzman tarafından "yeterli değil/eksik" olarak değerlendirildi. Altı (%22,2) yanıt ise iki uzman tarafından "uygun", bir uzman tarafından "uygun değil" olarak değerlendirildi. Genel olarak bakıldığında en az iki uzmanın "uygun", olarak nitelendirdiği toplamda 16 yanıt (%59,25) mevcuttu.

Sonuç: ChatGPT gibi YZ tabanlı sohbet programları hastaların medikal tavsiyeler aldığı bir uzmanın yerini alması şu aşamada mümkün görülmemektedir.

Anahtar Kelimeler: ChatGPT, tiroid kanseri, yapay zeka

Introduction

Large language models (LLMs) are AI products that use deep learning techniques, such as artificial neural networks, to replicate human language processing capabilities⁽¹⁾. They are capable of learning and processing vast amounts of language data from various sources. ChatGPT, developed by OpenAI (OpenAI, L.L.C., San Francisco, CA, USA) as a non-profit initiative and released on November 30, 2022, is an advanced chatbot that uses a text interface to understand and generate responses⁽²⁾. The accessibility of this application at no cost has encouraged people to use it as a tool for acquiring information on various topics, including health.

The operating principle of the application involves scanning all data on the internet using keywords from the queried questions. While the internet contains much accurate information, it also hosts incorrect or misleading data. Thus, the application can present erroneous information alongside accurate data. In sensitive topics such as health, especially concerning diseases like cancer, if the application provides incorrect information, it could cause medical and psychological distress to a patient^(3,4). Consequently, it is necessary for the responses given by the application to be evaluated by expert clinicians.

According to data from the World Health Organization, thyroid cancer is the second most common cancer among women in our country, a region endemic for goitre. It follows breast cancer, with approximately 13,500 cases annually (representing 5.9% of all cancers)⁽⁵⁾. Given this prevalence, there is a clear need for the public to access information in Turkish about thyroid cancer, both from the internet and through the ChatGPT application. Our study aims to evaluate the appropriateness of the ChatGPT application in

answering basic questions about thyroid cancer posed by non-healthcare professionals in Turkish.

Materials and Methods

A total of 27 Turkish questions were prepared in four different sections (general information, diagnosis, treatment and follow-up), which non-healthcare professionals could ask about thyroid cancer (Table 1). These questions were prompted twice in the free public version of ChatGPT 3.5 for consistency. The responses obtained were compiled using a survey via Google Forms. Three expert academicians in endocrine surgery, each holding a European Board of Endocrine Surgery certification and having a Web of Science H-index above 10, were asked to evaluate the responses provided by ChatGPT to the prepared Turkish questions on thyroid cancer.

Ethical approval was obtained for the study from the Ege University Medical Research Ethics Committee (decision no: 23-12T/32, date: 14.12.2023). An informed consent form was presented online to the three academicians, who provided their consent.

Statistical Analysis

The academicians categorized ChatGPT's responses into three groups: Appropriate, Inappropriate, and Insufficient/Incomplete. The responses were recorded using Microsoft 365 Excel. Descriptive datasets were compiled from an excel spreadsheet for the ChatGPT answers to each question.

Results

When evaluating the responses provided by ChatGPT to the prepared questions across four groups, of 27 responses: 9 (33.3%) were considered "appropriate" by two of the three experts and "insufficient/incomplete" by one expert. Six

responses (22.2%) were deemed “appropriate” by two experts and “inappropriate” by one expert. There were six responses (22.2%) categorized as “Insufficient/Incomplete” by two experts and “appropriate” by one. One response was judged as “appropriate” by all three experts, while another received a rating of 3 for “insufficient-incomplete”. One response was assessed differently by each of the three experts (Table 2).

The number of responses found “appropriate” by at least two experts was 16 (59.2%), while the responses considered “inappropriate” by at least two experts were only 2 (7.4%). Responses labelled as “insufficient/incomplete” by at least two experts totalled 8 (29.6%).

The response to the question “My ultrasound results mention “EU-TIRADS 4”. What does this mean?” under the Diagnosis section was evaluated as “appropriate” by all three experts (Figure 1). The response to “Who performs thyroid cancer surgery?” under the Treatment section was assessed differently by each expert (Figure 2). The response to “How is thyroid cancer surgery performed?” also under

the Treatment section was rated as “insufficient/incomplete” by all three experts (Figure 3).

Discussion

When existing questions were asked to traditional search engines, advertisements were typically encountered within the first one or two links. Subsequently, it was observed that the information obtained by clicking on links other than these was often hosted on health-related websites that generally lacked proper citations of sources. Accessing accurate information through traditional search engines was found to be significantly more challenging and time-consuming compared to ChatGPT. While ChatGPT is a highly effective and efficient artificial intelligence chat program, especially in scientific research and healthcare professional training due to its ability to quickly access vast information in various languages, there are still ethical issues associated with its use at the community level in addressing health problems. These include biases in data and concerns over the privacy and security of personal data⁽⁶⁾.

Table 1. The question's asked to ChatGPT in the four groups

General information	Diagnosis	Treatment	Follow-up
1- What is thyroid cancer?	1- What are the symptoms of thyroid cancer?	1- How is thyroid cancer treated?	1- How is follow-up done after thyroid cancer surgery?
2- How common is thyroid cancer?	2- How can I understand if I have thyroid cancer?	2- Who performs thyroid cancer surgery?	2- Does thyroid cancer recur after surgery?
3- What factors increase the risk of thyroid cancer?	3- How is thyroid cancer diagnosed?	3- Does thyroid cancer heal with surgery?	3- When I will recover and return to work after thyroid cancer surgery?
4- Is thyroid cancer genetic?	4- I have a lump in my neck, do I have thyroid cancer?	4- How is thyroid cancer surgery performed?	What happens if I don't take my thyroid medications?
5- Is thyroid cancer fatal?	5- My ultrasound result reported as “EU-TIRADS 4”. What does this mean?	5- What are the risks of thyroid cancer surgery?	
6- How can I prevent thyroid cancer?	6- Why do some thyroid cancer patients have hoarseness?	6- Do all thyroid cancer patients receive radiation therapy?	
7- Should I undergo screening for early detection of thyroid cancer?		7- How is radiation therapy performed in thyroid cancer?	
		8- How is treatment done for thyroid cancer that has spread to lymph nodes?	
		9- Why does hoarseness occur after thyroid surgery?	
		10- Is there any treatment for thyroid cancer other than surgery?	

Table 2. The appropriateness of the responses provided by ChatGPT to the prepared questions			
The number of experts evaluating the responses			The responses provided by ChatGPT to the questions
Appropriate	Inappropriate	Insufficient/incomplete	
3	-	-	1
-	3	-	-
-	-	3	1
2	1	-	6
2	-	1	9
1	2	-	1
1	-	2	1
1	1	1	1
-	2	1	1
-	1	2	6
Total			27

5- The ultrasound report mentions the term ‘EU-TIRADS 4’. What does this mean?

ChatGPT 3.5 Answer: "EU-TIRADS 4" refers to the *European Thyroid Imaging Reporting and Data System* (EU-TIRADS), which is a classification system used in the European Thyroid Imaging Reporting and Data System. This system is used to evaluate thyroid nodules and to determine their risk levels. It is designed to assist in determining the likelihood of malignancy in thyroid nodules and to guide clinical management accordingly.

The EU-TIRADS classification generally includes 5 categories, from 1 to 5. Here are the meanings of each category:

- 1. EU-TIRADS 1: Definitely benign nodule (non-cancerous).
- 2. EU-TIRADS 2: Probably benign nodule.
- 3. EU-TIRADS 3: Indeterminate nodule, with a higher probability of being benign.
- 4. EU-TIRADS 4: Suspicious nodule, likely benign but follow-up and possibly biopsy may be needed.
- 5. EU-TIRADS 5: Probably malignant nodule (high cancer risk), biopsy is recommended.

Therefore, a nodule classified as EU-TIRADS 4 is likely benign, but due to certain suspicious features, further evaluation and possibly a biopsy may be necessary. In this case, your doctor will likely recommend further evaluation, including fine needle aspiration biopsy and other tests. These tests help to determine whether the nodule is benign or malignant and assess the potential risk of cancer.

Figure 1. The question was evaluated as "Appropriate" by all three experts (the question asked of ChatGPT in Turkish is translated into English in this figure)

In the increasingly consumer-focused model of healthcare services, the unprecedented access to information could extend to patients using ChatGPT to form opinions on medical questions. A recent study revealed that 89% of people in the United States consult Google for health symptoms before visiting a doctor⁽⁷⁾. The popularity of the LLM-based AI chat

program ChatGPT has grown significantly over the past two years, demonstrating its potential use by patients as an access tool for health information. However, although the AI chat program ChatGPT can simplify our lives in many ways, its responses to Turkish questions about thyroid cancer, which requires specific expertise, were found to be

2- Who performs thyroid cancer surgery?

ChatGPT 3.5 Answer: Thyroid cancer surgery is usually performed by a surgeon, typically a head and neck surgeon or an endocrine surgeon. These types of surgeons are specialized professionals with expertise and experience in the thyroid gland and surrounding tissues.

Thyroid surgery can range from simple procedures such as removing thyroid nodules, to more complex operations such as the complete removal of the thyroid gland in patients with thyroid cancer. These procedures may include lobectomy (removal of one lobe), total thyroidectomy (removal of the entire thyroid gland), and if necessary, removal of nearby lymph nodes (neck dissection).

The extent of the surgery depends on several factors including the type of thyroid cancer, the size of the tumor, its spread, and the general health status of the patient. Surgery is generally part of the treatment for thyroid cancer and aims to remove the cancerous cells from the body.

Thyroid cancer surgeries are typically performed by experienced surgeons. After surgery, it is important to monitor the patient's recovery and hormone levels. Treatment may also include radioactive iodine therapy, hormone replacement therapy, or other methods. Patients are followed up and their treatment plans are adjusted as necessary by specialists such as endocrinologists (hormone doctors) or oncologists (cancer doctors).

Figure 2. The question was assessed differently by each expert (the question asked of ChatGPT in Turkish is translated into English in this figure)

appropriate by at least two experts in about 60% of the cases. On the other hand, a study by Zalzal et al.⁽⁸⁾ reported that responses provided by ChatGPT to patient questions about ENT diseases were quite satisfactory. Furthermore, a study examining responses by ChatGPT to general questions found that approximately 87% of responses to 38 questions about colon cancer management were deemed appropriate and consistent by at least two independent experts⁽⁹⁾. Another study evaluating the appropriateness of cardiovascular disease prevention recommendations provided by ChatGPT found that 84% of the recommendations were consistent and appropriate⁽¹⁰⁾.

Köroğlu et al.⁽¹¹⁾ found the responses given by ChatGPT on the management of thyroid nodules to be mostly accurate and reliable when evaluated by two expert endocrinologists. However, they did not consider it appropriate as a primary source for physicians, suggesting it could guide patients. The lower appropriateness rates in our study compared to others might be due to our use of the Turkish language. The primary mechanism of the ChatGPT program is to analyze data sources on the internet to generate answers. The lesser availability of Turkish sources compared to sources in more frequently used languages like English could also be a factor in the inadequacy of responses⁽¹²⁾. Deiana et al.⁽¹³⁾ study on

myths about vaccination asked questions of both ChatGPT 3.5 and ChatGPT 4.0, finding a 17% difference in clarity between the versions, which could be a factor contributing to the unsuitability of responses in our study. The paid 4.0 version, containing improvements and updates, is likely to provide more detailed and acceptable responses than the free version used in our study.

Some studies have shown that ChatGPT has similar accuracy rates in different languages⁽¹⁴⁾, but no literature reviews have encountered a study comparing Turkish with other languages. Future studies should consider this and design their studies to include comparisons between Turkish and other languages to achieve more effective results. A study on thyroid nodules by Campbell et al.⁽¹⁵⁾ found the accuracy rate of responses given by the ChatGPT application to be 69.2%, which aligns with our study, suggesting that lower accuracy rates in two studies could be topic-dependent. The appropriateness of responses given by ChatGPT on thyroid nodules and treatment may be lower than other health topics. This deficiency could be addressed over time by making more databases available to applications like ChatGPT. The readability level of responses provided by ChatGPT for community use has been indicated by international standards⁽¹⁶⁾ to be higher than middle school

4- How is thyroid cancer surgery performed?

ChatGPT 3.5 Answer: Thyroid cancer surgery involves the complete or partial removal of the thyroid gland and is usually performed by head and neck surgeons or endocrine surgeons. The surgical procedure may vary depending on the type, size, and stage of the thyroid cancer. Below are the general steps involved in thyroid cancer surgery:

1. Preparation and Anesthesia:
 - The patient's overall health status is evaluated before the operation.
 - The anesthesiologist administers general anesthesia to ensure the patient sleeps through the surgery and feels no pain.
2. Determining the Incision Site:
 - The surgeon makes an incision, usually in the neck area, to access the thyroid gland.
 - The location of the incision may vary depending on the type of surgery to be performed (e.g., thyroid lobectomy or total thyroidectomy), the surgeon's preferences, and the patient's condition.
3. Removal of the Thyroid Gland:
 - The surgeon accesses the thyroid gland through the incision.
 - Depending on the extent of the disease, type of cancer, and tumor size, the operation may involve removal of one lobe (lobectomy), the entire thyroid (total thyroidectomy), or part of the thyroid.
 - If needed, surrounding tissues or lymph nodes may also be removed.
4. Closure:
 - After removing the thyroid gland, the incision site is closed with sutures.
 - A surgical drain may be placed if needed.
5. Postoperative Monitoring and Care:
 - The patient is monitored after surgery, and their recovery is assessed.
 - Thyroid hormone levels are checked and managed accordingly.
 - Radioactive iodine therapy or other treatments may be required, depending on the pathology results.

In some cases, patients may be discharged the same day, while others may need to stay in the hospital for a short period. After discharge, regular follow-up visits are scheduled to monitor hormone levels, ensure healing, and determine if further treatment is needed.

Figure 3. The question was assessed as “insufficient/incomplete” by all three experts (the question asked of ChatGPT in Turkish is translated into English in this figure)

level, which shows a gap in reaching the public effectively in health education. No study evaluating the language level of responses in Turkish has been found, although language level assessments have been researched in English. Future study designs are advised to consider this issue.

ChatGPT, a 24/7 accessible application, has become an increasingly popular AI-based chat program for medical advice, yet it must be remembered that ChatGPT is not a medical application. If used without proper oversight in the health sector, it is expected to lead to medical and paramedical problems.

Study Limitations

This study has limitations such as the limited access to accurate data of ChatGPT due to the fact that the questions are in Turkish but most of the resources accessible on the internet are in English, the small number of academics who check the questions and meet the selection criteria, and the inability to evaluate the access to ChatGPT 4.0 (paid version).

Conclusion

Although technological advancements are increasingly integrating artificial intelligence into our daily lives, and its usage by the public is growing, it is currently not deemed appropriate for AI-based chat programs like ChatGPT to replace medical professionals. Such programs should not provide advice on specific issues requiring professional health services, such as thyroid cancer. Certainly, as this AI program continues to evolve, it is expected to greatly benefit the health sector by providing doctors with the opportunity to save time in clinics, thereby reaching more patients effectively and offering patients 24/7 access to information. However, patients should obtain the most reliable and accurate information about their conditions through specialists in the field.

Ethics

Ethics Committee Approval: Ethical approval was obtained for the study from the Ege University Medical Research Ethics Committee (decision no: 23-12T/32, date: 14.12.2023).

Informed Consent: An informed consent form was presented online to the three academicians, who provided their consent.

Footnotes

Authorship Contributions

Surgical and Medical Practices: Y.T., B.E.B., Concept: R.G.İ., Design: Y.T., Data Collection or Processing: B.E.B., Analysis or Interpretation: B.E.B., A.C.D., S.T., Ö.M., Literature Search: B.E.B., M.Ö., Writing: Y.T., B.E.B.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

References

1. Muehmel K. What is a large language model, the tech behind ChatGPT? Available from: <https://blog.dataiku.com/large-language-model-chatgpt>
2. OpenAI. Introducing ChatGPT. Available from: <https://openai.com/blog/chatgpt>
3. Başkale HA, Serçekuş P, Partlak Güneşen N. Investigation of cancer patients' information sources, information needs and expectations of health professionals. *J Psy Nurs*. 2015;6:65-70.
4. Liebl P, Seilacher E, Koester MJ, Stellamanns J, Zell J, Hübner J. What cancer patients find in the internet: the visibility of evidence-based patient information - analysis of information on German websites. *Oncol Res Treat*. 2015;38:212-8.
5. International Agency for Research on Cancer. Available from: <https://gco.iarc.fr/today/data/factsheets/populations/792-turkey-fact-sheets.pdf>
6. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)*. 2023;11:887.
7. Dubin JA, Bains SS, Chen Z, et al. Using a Google web search analysis to assess the utility of ChatGPT in total joint arthroplasty. *J Arthroplasty*. 2023;38:1195-202.
8. Zalzal HG, Abraham A, Cheng J, Shah RK. Can ChatGPT help patients answer their otolaryngology questions? *Laryngoscope Investig Otolaryngol*. 2024;9:e1193.
9. Emile SH, Horesh N, Freund M, et al. How appropriate are answers of online chat-based artificial intelligence (ChatGPT) to common questions on colon cancer? *Surgery*. 2023;174:1273-5.
10. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. *JAMA*. 2023;329:842-4.
11. 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.
12. Sarker IH. AI-based modeling: techniques, applications and research issues towards automation, intelligent and smart systems. *SN Comput Sci*. 2022;3:158.
13. Deiana G, Dettori M, Arghittu A, Azara A, Gabutti G, Castiglia P. Artificial intelligence and public health: evaluating ChatGPT responses to vaccination myths and misconceptions. *Vaccines (Basel)*. 2023;11:1217.
14. Shao CY, Li H, Liu XL, et al. Appropriateness and comprehensiveness of using ChatGPT for perioperative patient education in thoracic surgery in different language contexts: survey study. *Interact J Med Res*. 2023;12:e46900.
15. Campbell DJ, Estephan LE, Sina EM, et al. Evaluating ChatGPT responses on thyroid nodules for patient education. *Thyroid*. 2024;34:371-7.
16. Rooney MK, Santiago G, Perni S, et al. Readability of patient education materials from high-impact medical journals: a 20-year analysis. *J Patient Exp*. 2021;8:237437352199884.