

A Novel Approach: Evaluating ChatGPT's Utility for the Management of Thyroid Nodules

Review began 10/13/2023
Review ended 10/19/2023
Published 10/24/2023

© Copyright 2023

Köroğlu et al. This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Ekin Y. Köroğlu¹, Sevgül Faki¹, Nagihan Beştepe¹, Abbas A. Tam², Neslihan Çuhacı Seyrek², Oya Topaloglu², Reyhan Ersoy², Bekir Cakir²

1. Endocrinology and Metabolism, Ankara City Hospital, Ankara, TUR 2. Endocrinology and Metabolism, Ankara Yıldırım Beyazıt University School of Medicine, Ankara, TUR

Corresponding author: Ekin Y. Köroğlu, eyigitkoroglu@hotmail.com

Abstract

Background and objective

Artificial intelligence (AI) applications such as Chat Generative Pre-Trained Transformer (ChatGPT) created by OpenAI, which represent the revolutionary aspects of today's technology, have benefitted professionals in many fields and society at large. In this study, we aimed to assess how effective is ChatGPT in helping both the patient and the physician manage thyroid nodules, a very common pathology.

Methods

Fifty-five questions frequently asked by patients were identified and asked to ChatGPT. Subsequently, three cases of thyroid nodules were progressively presented to ChatGPT. The answers to patient questions were scored for correctness and reliability by two endocrinologists. As for the cases, diagnostic and therapeutic approaches provided by ChatGPT were analyzed and scored by two endocrinologists for correctness, safety, and usability. The responses were evaluated by using 7-point Likert-type scales designed by us.

Results

The answers to patient questions were found to be mostly correct and reliable by both raters (Rater #1: 6.47 ± 0.50 and 6.27 ± 0.52; Rater #2: 6.18 ± 0.92 and 6.09 ± 0.96). Regarding the management of cases, ChatGPT's approach was found to be largely correct, safe, and usable by Rater #1, while Rater #2 evaluated the approaches as partially or mostly correct, safe, and usable.

Conclusion

Based on our findings, ChatGPT can be used as an informative and reliable resource for managing patients with thyroid nodules. While it is not suitable to be used as a primary resource for physicians, it has the potential to be a helpful and supportive tool.

Categories: Endocrinology/Diabetes/Metabolism, Internal Medicine, Healthcare Technology

Keywords: chatgpt, thyroid, thyroid cancer, artificial intelligence, thyroid nodule

Introduction

Recent years have witnessed the rapid development of artificial intelligence (AI) applications that assist in accessing and using information rationally on the web. Chat Generative Pre-Trained Transformer (ChatGPT), created by OpenAI and made available in November 2022, has become an easily accessible resource for both professionals in various fields and the community at large [1]. Many studies have been conducted on the use of this resource in medicine. These studies have tested the adequacy, usability, and reliability of ChatGPT for both patients and healthcare providers [2-4].

In a study on the use of ChatGPT for treating solid tumors, the tool was able to respond in a manner that was partially consistent with the National Comprehensive Cancer Network (NCCN) guidelines and was found to be open to improvement [2]. In a study on the treatment of thumb arthritis from the perspective of plastic surgery, ChatGPT was determined to be effective in accessing information but insufficient in producing solutions [3].

Thyroid nodules have a prevalence of 60% among the global population. Approximately 5% of these patients have malignant features [5]. Ultrasonography and fine-needle aspiration biopsy (FNAB) are commonly used to determine the malignant potential of thyroid nodules [6]. National and international guidelines (e.g., American Thyroid Association guidelines) guide healthcare professionals about the use of these methods in specific patients and situations, as well as how to evaluate the results [7]. These guidelines are updated in the wake of advancements in clinical experience and new findings regarding thyroid nodules and cancers, such as the recent update of the histopathological classification of thyroid cancers [8].

How to cite this article

Köroğlu E Y, Faki S, Beştepe N, et al. (October 24, 2023) A Novel Approach: Evaluating ChatGPT's Utility for the Management of Thyroid Nodules. Cureus 15(10): e47576. DOI 10.7759/cureus.47576

As AI applications are easily accessible, they can also act as consultants for patients in medical matters [1]. Thyroid nodules are one of the most frequently discussed medical conditions by patients because of their high prevalence. These AI applications, which are now widely used, are models that can be developed for use in medical practice. Therefore, it is crucial to test these applications to determine how usable and reliable they are for medical practitioners [2]. Although varying results have been obtained in the studies conducted so far, the consensus among many is that practices should be developed in this area [2-4].

This is the first study on the usability of AI applications in the management of thyroid modules. The widespread use of ChatGPT and other similar AI applications warrants the need to test their reliability and usability. Therefore, in this study, we aimed to assess how accurate and reliable is ChatGPT in helping patients and doctors in the management of thyroid modules.

Materials And Methods

The study consisted of two parts. The first part involved collecting 55 questions that are frequently asked by patients in daily endocrinology practice regarding thyroid nodules (Table 1).

| Questions | |
|----------------------------|--|
| General information | |
| Q-1 | What is a thyroid nodule? |
| Q-2 | Why does a thyroid nodule occur? |
| Q-3 | Is a thyroid nodule genetic? |
| Q-4 | Is a thyroid nodule dangerous? |
| Q-5 | How to detect the presence of a thyroid nodule? |
| Q-6 | Does the thyroid nodule disrupt the functioning of the thyroid gland? |
| Q-7 | Is it more dangerous to have more than one nodule? |
| Q-8 | Does the thyroid's slow or fast functioning increase the risk of cancer? |
| Q-9 | Is thyroid cancer deadly? |
| Q-10 | Does thyroid cancer spread throughout the body? |
| Diagnosis process | |
| Q-11 | How is a thyroid nodule examined? |
| Q-12 | Is the presence of a thyroid nodule detected by blood tests? |
| Q-13 | How to understand whether a thyroid nodule is dangerous? |
| Q-14 | Does the ultrasound definitively show whether the nodule is good or bad? |
| Q-15 | Do all thyroid nodules need to be biopsied? |
| Q-16 | Can it be understood that the nodule is good or bad without taking a biopsy? |
| Q-17 | Can scintigraphy be used instead of biopsy? |
| Q-18 | What are the risks of a biopsy? |
| Q-19 | Is a biopsy a painful procedure? |
| Q-20 | Is a biopsy a surgery-like procedure? |
| Q-21 | If the nodule is malignant, will the cancerous cells spread while the biopsy is taken? |
| Q-22 | Does the biopsy give definitive results? |
| Q-23 | What does the non-diagnostic result mean? What is the cancer risk? |
| Q-24 | When should a biopsy be performed again after a non-diagnostic result? |
| Q-25 | If the nodule is malignant, is waiting until the repeat biopsy time risky? |
| Q-26 | What does atypia of undetermined significance mean? What is the cancer risk? |

| | |
|--------------------------|--|
| Q-27 | What does high calcitonin mean? |
| Treatment process | |
| Q-28 | Should all thyroid nodules be treated? |
| Q-29 | What are the treatment options for thyroid nodules? |
| Q-30 | What needs to be eaten to make the nodules disappear? |
| Q-31 | Is there a drug that shrinks the thyroid nodule? |
| Q-32 | Is surgery the only treatment option for a nodule thought to be malignant? |
| Q-33 | Is it possible to get rid of thyroid cancer with surgery? |
| Q-34 | What are the risks of thyroid surgery? |
| Q-35 | What happens if surgery is not performed on a nodule that is thought to be malignant? |
| Q-36 | If the postoperative pathology result is cancer, is chemotherapy given? |
| Q-37 | Will there be a need for another operation after the first operation? |
| Q-38 | Is it risky to have surgery for the second time? |
| Q-39 | What is radioactive iodine? Why is it given? |
| Q-40 | Do all thyroid cancer patients receive radioactive iodine treatment? |
| Q-41 | Is radioactive iodine therapy dangerous? Does it cause cancer? |
| Q-42 | Will radioactive iodine harm people near the patient? |
| Q-43 | How is radioactive iodine treatment given while breastfeeding? |
| Q-44 | What happens if radioactive iodine treatment is not taken although it is recommended? |
| Q-45 | Is it possible to give radioactive iodine for the second time? |
| Follow-up process | |
| Q-46 | How often should an ultrasound be performed on a benign nodule? |
| Q-47 | Can a benign nodule become malignant in the future? |
| Q-48 | What is the probability of disease recurrence after thyroid cancer treatment? |
| Q-49 | How often is a follow-up required after thyroid cancer treatment? |
| Q-50 | How does the disease relapse during follow-up? |
| Q-51 | How can it be understood that the disease has recurred in the follow-up? |
| Q-52 | Is it necessary to take the levothyroxine sodium drug given after the treatment? What if it is not received? |
| Q-53 | How often should blood tests be conducted in thyroid cancer follow-up? |
| Q-54 | How often should scintigraphy be done in thyroid cancer follow-up? |
| Q-55 | Is there any food that should not be eaten during the follow-up of thyroid nodules and thyroid cancer? |

TABLE 1: Frequently Asked Questions by Patients Regarding Thyroid Nodules

Q: question

The questions were prepared based on data from interviews conducted during the daily outpatient examination and Google Trends. We asked 10 endocrinologists working in our clinic to list the 10 questions most frequently asked by patients about thyroid nodules. Among the 100 questions obtained, the 50 most frequently repeated questions and five questions obtained from Google Trends were included in the study. The included questions were validated by an expert panel and classified into four main headings: general information (10 questions), diagnosis (17 questions), treatment (18 questions), and follow-up (10 questions). The questions were entered into the chat section of the ChatGPT chatbot, and the answers were recorded. The version of ChatGPT dated June 17, 2023, was used in this study. The answers given to these

questions by ChatGPT were evaluated by two independent endocrinologists. For evaluation, correctness and reliability criteria were designed based on Likert-type scales, with scores ranging from 1 to 7, with higher scores indicating a higher level of correctness and reliability (Table 2).

| Score | Correctness | Reliability |
|-------|----------------------------------|---|
| 7 | Completely correct information | The answer guides the patient completely correctly |
| 6 | Mostly correct information | The answer guides the patient mostly correctly |
| 5 | Partially correct information | The answer guides the patient partially correctly |
| 4 | Insufficient information | The answer does not guide the patient in a right or wrong way |
| 3 | Partially incorrect information | The answer partially misleads the patient |
| 2 | Mostly incorrect information | The answer mostly misleads the patient |
| 1 | Completely incorrect information | The answer completely misleads the patient |

TABLE 2: Correctness and Reliability Scales Used by Raters to Assess ChatGPT's Answers to the Patient Questions

We designed the scoring scales ourselves given the lack of similar studies in the literature, as well as the lack of a standardized scale.

In the second part of the study, the details of three patients who were seen in the endocrinology clinic were presented to ChatGPT pertaining to a certain setting along with relevant medical terminology, without providing any information on the identity of the patients. The cases were determined by an expert panel as those with a benign thyroid nodule, toxic multinodular goiter, and a malignant thyroid nodule, which are frequently seen in endocrinology practice. We planned to use three different cases in order to evaluate the responses to different clinical situations by ChatGPT. When the cases were presented to ChatGPT, it was ensured the real data was presented without any changes and in a way that did not disrupt the question flow. Informed consent was obtained from the patients. The cases were entered into the ChatGPT chat section and they included information on age, sex, and complaints the patients presented with at the outpatient clinic. After answering the questions from ChatGPT regarding the anamnesis, we asked ChatGPT to schedule the tests and examinations. The test results were then submitted to the ChatGPT chatbot without reference intervals or medical interpretation. ChatGPT was asked to draw up a treatment and follow-up plan based on the information provided. The responses of ChatGPT at each stage were evaluated by two independent endocrinologists. Correctness, safety, and usability criteria designed by us based on Likert-type scales, with scores ranging from 1 to 7, were used for evaluation (Table 3).

| Score | Correctness | Safety | Usability |
|-------|--|--|---|
| 7 | Completely correct recommendation | Completely beneficial recommendation | All recommendations can be used in practice |
| 6 | Mostly true recommendation, no false recommendation | Mostly beneficial recommendation, not harmful | Most of the recommendations can be used in practice |
| 5 | Partially true recommendation, no false suggestion | Partially beneficial recommendation, not harmful | Recommendations can be partially used in practice |
| 4 | Ineffective recommendation for the diagnosis-treatment process | The recommendation does not affect the patient | Recommendations not usable in practice |
| 3 | Partially false recommendation | Partially harmful | Recommendations are better not used in practice |
| 2 | Mostly false recommendation | Mostly harmful recommendation | Recommendations should not be used in practice |
| 1 | Completely false recommendation | Completely harmful recommendation | Recommendations should never be used in practice |

TABLE 3: Correctness, Safety, and Usability Criteria Used by Raters to Evaluate ChatGPT's Approaches Toward the Cases

This was an observational study, and the Ankara City Hospital Research Ethics Committee confirmed that no ethical approval was required.

The IBM SPSS Statistics software package version 25.0 (IBM Corp., Armonk, NY) was used for statistical analysis. The Shapiro-Wilk test was used to determine whether or not variables exhibited a normal distribution. We presented the descriptive statistics as the median (minimum-maximum) for non-normally distributed variables and mean \pm standard deviation (SD) for normally distributed variables. Student's t-test was used for parametric variables and the Mann-Whitney U test was used for nonparametric variables to investigate the differences between groups. A p-value <0.05 was considered statistically significant.

Results

All 55 questions frequently asked by patients were answered by ChatGPT. The mean values of the scores given by the raters to the answers are compared in Table 4.

| | General information | Diagnostic process | Treatment process | Follow-up process | Total |
|---------------|---------------------|--------------------|-------------------|-------------------|-----------------|
| Rater #1 | | | | | |
| Correctness | 6.20 \pm 0.42 | 6.64 \pm 0.49 | 6.61 \pm 0.50 | 6.20 \pm 0.42 | 6.47 \pm 0.50 |
| Reliability | 5.90 \pm 0.56 | 6.58 \pm 0.50 | 6.27 \pm 0.46 | 6.10 \pm 0.31 | 6.27 \pm 0.52 |
| Rater #2 | | | | | |
| Correctness | 6.30 \pm 0.67 | 6.29 \pm 0.77 | 6.38 \pm 0.97 | 5.50 \pm 1.08 | 6.18 \pm 0.92 |
| Reliability | 6.30 \pm 0.82 | 6.11 \pm 0.99 | 6.22 \pm 0.87 | 5.60 \pm 1.17 | 6.09 \pm 0.96 |
| Correctness p | 0.588 | 0.179 | 0.646 | 0.096 | 0.185 |
| Reliability p | 0.188 | 0.167 | 0.788 | 0.298 | 0.718 |

TABLE 4: Scores Given by the Raters to ChatGPT's Answers to Patient Questions

The data is represented as mean \pm SD

SD: standard deviation

The mean correctness and reliability scores given by the raters with regard to general information, diagnostic process, treatment process, and follow-up process of thyroid nodules were similar. The lowest correctness and reliability scores given by Rater #1 were 6 and 5, respectively. Rater #2 gave 3 points as the lowest correctness score and 4 points as the lowest reliability score. The answer to question 29, which received a correctness score of 3 and a reliability score of 4 by Rater #2, is shown in Figure 1.

EY

What are the treatment options for thyroid nodules?



The treatment options for thyroid nodules depend on various factors, including the characteristics of the nodule, the presence of symptoms, the risk of malignancy, and individual patient factors. Here are some common treatment options:

1. **Observation/Watchful Waiting:** If the thyroid nodule is small, benign, and not causing significant symptoms or health concerns, a conservative approach called observation or watchful waiting may be recommended. Regular monitoring through periodic clinical examinations and imaging studies is conducted to assess any changes in size or appearance of the nodule over time.
2. **Thyroid Hormone Suppression:** In some cases, especially if the thyroid nodule is causing hyperthyroidism or if there are multiple nodules, thyroid hormone medication may be prescribed. The medication aims to suppress the production of thyroid-stimulating hormone (TSH), which can help reduce the size or activity of the nodules.
3. **Fine-Needle Aspiration Biopsy (FNAB):** If the nodule is suspicious or has concerning features, a fine-needle aspiration biopsy (FNAB) may be performed to obtain a sample of cells for pathological evaluation. The results of the biopsy help guide further treatment decisions.
4. **Surgical Treatment:** Surgery may be recommended for various reasons, including:
 - **Malignant nodules:** If the biopsy confirms thyroid cancer, surgical removal of the affected thyroid tissue may be necessary. The extent of surgery will depend on the type and stage of cancer.
 - **Large or symptomatic nodules:** If the nodule is causing significant symptoms, such as difficulty swallowing, breathing problems, or cosmetic concerns, surgical removal (partial or total thyroidectomy) may be considered.
 - **Indeterminate or suspicious nodules:** If the biopsy results are inconclusive (e.g., atypia, follicular lesion), surgery may be recommended to obtain a definitive diagnosis and assess the extent of disease.
 - **Non-malignant but problematic nodules:** In some cases, non-malignant nodules that cause significant symptoms or cosmetic issues may be surgically removed.
5. **Radiofrequency Ablation (RFA):** RFA is a minimally invasive procedure where heat energy is used to destroy the thyroid nodule. It may be an option for selected cases, such as benign solid nodules or recurrent cystic nodules.
6. **Ethanol Ablation:** Ethanol ablation involves injecting ethanol (alcohol) into a thyroid cyst or certain types of nodules to shrink or eliminate them. It is generally used for cystic nodules that reoccur after repeated drainage or for patients who are not surgical candidates.

The choice of treatment depends on factors such as nodule size, characteristics, symptoms, risk of malignancy, patient preference, and expert medical advice. It's important to consult with an endocrinologist or a thyroid specialist who can evaluate your specific case and guide you on the most appropriate treatment option based on your individual circumstances.

FIGURE 1: Answer to Question 29 Given by ChatGPT

Rater # 2 stated that they gave these scores to this answer because ChatGPT recommended thyroid hormone suppression for the treatment of thyroid nodules.

Case 1 involved a 24-year-old female musician who had had a thyroid nodule for three years and complained of swelling in the neck; her details were presented to the ChatGPT chatbot. Based on the questions asked by ChatGPT and the examinations requested, information about the patient was provided to ChatGPT, as shown

in Figure 2.

EY

Patient's TSH:3.22 nIU/ml, fT4:1.11 ng/dl, fT3:3.48 pg/mL, anti-tpo:33 U/mL, anti-tg:<1.3 IU/mL, Tg : 95.6 ng/mL. Liver functions, kidney functions, electrolytes and complete blood count were normal. As a result of scintigraphy, the thyroid gland was evaluated as normal functioning. Thyroid ultrasound revealed a 15×28×42 mm well-circumscribed isoechoic nodule in the right lobe. What should be the next step in patient management?



Based on the provided test results and imaging findings, the next step in patient management would depend on several factors, including the size of the nodule, its characteristics, and the patient's clinical presentation. However, given the information provided, here are some potential next steps:

1. Repeat Fine Needle Aspiration Biopsy (FNAB): Since the previous FNAB was non-diagnostic, a repeat FNAB can be considered to obtain a more conclusive result. The size of the nodule warrants further evaluation to rule out malignancy or determine the nature of the nodule.
2. Consider a Core Needle Biopsy: If the repeat FNAB is again non-diagnostic or inconclusive, a core needle biopsy may be recommended. This type of biopsy uses a larger needle to obtain a larger tissue sample for more accurate evaluation.
3. Referral to an Endocrinologist or Thyroid Specialist: Given the complexity of the case and the need for further evaluation, it would be advisable to involve an endocrinologist or a specialist with expertise in thyroid disorders. They can provide a comprehensive evaluation, interpretation of results, and guide the management plan.
4. Surveillance and Monitoring: If the repeat FNAB or core needle biopsy does not indicate malignancy, close surveillance with periodic follow-up visits, thyroid ultrasound, and blood tests may be recommended to monitor the nodule's growth and any potential changes in its characteristics.
5. Surgical Consultation: Depending on the findings of further evaluation, such as high suspicion for malignancy or concerning features of the nodule, a surgical consultation may be warranted to discuss the option of surgical removal of the nodule or part of the thyroid gland.

The final decision regarding the next step in patient management should be made in consultation with a healthcare professional, taking into consideration the patient's individual circumstances, preferences, and the clinical judgment of the treating physician.

FIGURE 2: One of the Answers Given by ChatGPT During the Evaluation of Case 1

Then, the information that repeated FNAB results were non-diagnostic and that the patient had invasive procedural anxiety due to being a musician was presented to ChatGPT. Molecular testing, advanced thyroid imaging, or active follow-up were recommended by ChatGPT. Five questions were posed in total regarding the case. The correctness, safety, and usability scores given by Rater #1 for this case were 6.20 ± 0.44 , 6.20 ± 0.44 , and 6.40 ± 0.54 , while the scores given by Rater #2 were 5.00 ± 1.22 , 5.00 ± 1.22 , and 5.40 ± 1.51 , respectively. The lowest correctness, safety, and usability score that the raters gave was 3, which was assigned to the same question by the same rater. This question and the given answer are summarized above. We found no statistically significant difference between the raters' scores in terms of correctness ($p=0.257$), safety ($p=0.257$), and usability ($p=0.136$).

In Case 2, details of a 58-year-old male patient with palpitations and a thyroid-stimulating hormone (TSH) value of <0.008 nIU/mL were presented to ChatGPT. This case was selected as a case of toxic multinodular goiter (TMNG) with benign cytology of nodules. Based on 11 questions, the patient was diagnosed with TMNG, the nodules were benign, and anti-thyroid treatment was recommended by ChatGPT. The correctness, safety, and usability scores given by Rater #1 for this case were 6.09 ± 0.70 , 6.09 ± 0.53 , and 6.09 ± 0.53 , while the scores given by Rater #2 were 5.72 ± 1.19 , 5.72 ± 1.27 , and 5.90 ± 1.13 , respectively. We

found no statistically significant difference between the raters' scores in terms of correctness ($p=0.393$), safety ($p=0.393$), and usability ($p=0.637$). The lowest correctness, safety, and usability score the raters gave was 3, which was given to the same question by the same rater. This question and its answer are shown in Figure 3.



Two months after the start of the treatment, the patient's thyroid function tests were normal. FNAB result was seen as benign. What should be the next step?



If two months after starting methimazole treatment, the patient's thyroid function tests have normalized and the fine needle aspiration biopsy (FNAB) result indicates a benign finding, the next step would typically involve the following considerations:

1. Continue methimazole treatment: If the patient is currently on methimazole and has achieved normal thyroid function, it is generally recommended to continue the medication for a total treatment duration of 12 to 18 months. This duration helps ensure disease remission and stabilize thyroid function over the long term.
2. Regular follow-up and monitoring: The patient should be scheduled for regular follow-up visits with an endocrinologist or thyroid specialist to monitor thyroid function and overall health. Thyroid function tests, including levels of thyroid-stimulating hormone (TSH), free thyroxine (fT4), and free triiodothyronine (fT3), should be checked at appropriate intervals to assess the ongoing response to treatment and maintain thyroid function within the normal range.
3. Evaluate medication side effects: Methimazole can potentially cause side effects, such as rash, itching, or liver dysfunction. During follow-up visits, the patient's overall well-being and any potential adverse effects related to the medication should be assessed.
4. Patient education and counseling: It is important to educate the patient about the importance of medication adherence, potential signs of medication side effects, and the need for regular follow-up visits. Patient counseling should also include lifestyle modifications, such as stress management, a healthy diet, and avoidance of substances that may exacerbate hyperthyroidism symptoms (e.g., iodine-rich foods or supplements).
5. Discuss future management plans: Since the FNAB result was benign and the patient's thyroid function tests have normalized, a discussion about the long-term management plan should take place. This may involve considering the appropriate time for tapering off methimazole after completing the recommended treatment duration, as well as ongoing monitoring of thyroid function and nodules.

It is important to note that the next steps may vary depending on the patient's individual characteristics, preferences, and the clinical judgment of the healthcare provider. Close collaboration between the patient and the healthcare team is crucial in making informed decisions and optimizing the management of the patient's hyperthyroidism.

FIGURE 3: Answer with the Lowest Correctness, Safety, and Usability Score in Case 2

The raters stated that they gave these scores because ChatGPT recommended anti-thyroid treatment for 12-18 months, while surgery was not recommended.

In case 3, details of a 75-year-old female patient with an 8 x 9.8 x 12.4-mm hypoechoic nodule with microcalcifications, which was located in the left inferior part of the thyroid gland on thyroid ultrasound, were presented to ChatGPT. The case involved a patient whose FNAB result raised concerns about malignancy. The patient had one suspected lymphadenopathy on ultrasonography. The lymph node FNAB result was non-diagnostic, and thyroglobulin (Tg) washout was negative. The pathological diagnosis of the patient was invasive-encapsulated variant papillary thyroid cancer measuring 1.1 x 0.6 cm in diameter in the left lobe. Surgery was recommended and postoperative radioactive iodine treatment was administered to the patient.

ChatGPT requested an FNAB from the patient's suspicious nodule, as well as a neck ultrasound with the results. It then requested an FNAB from the suspected lymph node, but not a Tg washout. The results of the patient's lymph node FNAB performed twice were given as non-diagnostic. At this stage, we were asked whether or not the Tg washout could be performed, following which Tg washout was recommended. The serum Tg result was presented as 49 ng/mL and the Tg washout result was 0.1 ng/mL. ChatGPT recommended neck CT, 18-fluorodeoxyglucose positron emission tomography, and a third lymph node FNAB (Figure 4).

EY Thyroglobulin washing in the suspected lymph node was 0.1 ng/ml and simultaneous serum thyroglobulin level was 49 ng/ml. What is the next step?



If the thyroglobulin washing from the suspected lymph node was 0.1 ng/ml and the simultaneous serum thyroglobulin level was 49 ng/ml, there are a few possible interpretations and next steps to consider:

1. Confirm metastatic thyroid cancer: A significantly elevated serum thyroglobulin level suggests the possibility of metastatic thyroid cancer. The low thyroglobulin level in the lymph node washout may indicate incomplete or insufficient collection of thyroglobulin during the procedure. In this case, the next step would be to further investigate for metastatic disease to confirm the presence and location of metastases.
2. Repeat lymph node biopsy: Given the non-diagnostic results from previous biopsies, it may be necessary to repeat the lymph node biopsy to obtain a more representative sample. This can help provide a definitive diagnosis and guide further management decisions.
3. Imaging studies: Imaging studies, such as a neck ultrasound, computed tomography (CT) scan, or positron emission tomography (PET) scan, may be recommended to assess the extent of disease and identify potential sites of metastasis. These imaging modalities can help determine the presence of additional lymph nodes involved, distant metastases, or any other areas of concern.
4. Consultation with an endocrinologist or oncologist: It is crucial to consult with an endocrinologist or an oncologist who specializes in managing thyroid cancer to discuss the findings and determine the most appropriate course of action. They can evaluate the patient's overall clinical status, review the imaging and laboratory results, and guide further diagnostic and treatment decisions.

The next steps will depend on a comprehensive evaluation of the patient's clinical picture, including the histopathological results, imaging findings, and the overall clinical context. Consulting with a specialist experienced in managing thyroid cancer will provide the most

FIGURE 4: Answer with the Lowest Correctness, Safety, and Usability Score in Case 3

After the FNAB result was considered to be benign, thyroidectomy was recommended to the patient. ChatGPT stated that a risk classification should be made for the decision of radioactive iodine treatment based on the pathology result of the patient, but it did not provide a risk classification based on the available data.

The evaluation of Case 3 involved 17 questions. The correctness, safety, and usability scores given by Rater #1 for this case were 6.25 ± 0.56 , 6.11 ± 0.48 , and 6.17 ± 0.39 , while the scores given by Rater #2 were 5.25 ± 1.14 , 5.25 ± 1.03 , and 5.58 ± 1.17 , respectively. The lowest correctness and usability score was 2, and the lowest safety score was 3, both of which were given for the same answer (Figure 4). While there was a statistically significant difference in terms of correctness ($p=0.03$) and safety ($p=0.03$) scores of raters, there was no significant difference in usability scoring ($p=0.059$).

Throughout the entire study, no response from ChatGPT received a score of 1 for correctness, reliability, safety, or usability by the raters.

Discussion

Based on our findings, ChatGPT appears to be a mostly correct and reliable resource for patients. In clinical practice, although it cannot be solely relied upon for decision-making, it can be used as a supporting tool for physicians. In the present study, the evaluation of ChatGPT's answers to patient questions regarding thyroid nodules showed that the overall mean value of the correctness and reliability scores given by both raters was between 6.09 and 6.47. In other words, the recommendations given by ChatGPT to the patients are mostly correct and guide the patients correctly.

In the evaluation of Rater #1, the highest correctness and reliability scores were given to recommendations regarding the diagnostic process, while the lowest scores were given to answers related to general information and the follow-up process in correctness, and answers related to general information in reliability. Rater #2 gave the highest correctness scores to the answers related to the treatment process, while the highest reliability scores were given to the answers related to general information. The lowest correctness and reliability scores were given for recommendations related to the follow-up process. However, there was no significant difference between the evaluations of both raters in any particular category and overall.

In the evaluation of Case 1 with ChatGPT, the case was concluded with the shortest number of questions. In the evaluation of this case management by Rater #1, the mean scores were between 6.20 and 6.40 in terms of correctness, safety, and usability. In the evaluation by Rater #2, the scores were between 5.00 and 5.40, which were lower than those of Rater #1. However, there was no statistically significant difference between the two evaluations. In this case, the answer with the lowest scores among the ChatGPT answers is shown in Figure 2. This figure shows that ChatGPT can evaluate laboratory values and thyroid ultrasound findings even without reference values or medical interpretation. In addition, ChatGPT offers more than one option in terms of its recommendation. It stated that a healthcare professional should be consulted to decide which of these options should be used. This approach was evident in all of the answers and suggestions. This was judged by both the raters and us as a safety-enhancing approach.

In the evaluation of the responses to Case 2, the scores given by both raters were statistically similar to each other. The answers and recommendations evaluated by both raters were partially and mostly correct, safe, and usable. The answer with the lowest scores is shown in Figure 3. The raters considered the lack of surgical treatment options in this response as a deficiency. Again, in this answer, more than one recommendation was given and it was suggested that the decision should be made in consultation with a healthcare professional.

ChatGPT's assessment of Case 3 involved the highest number of questions among the cases. In the evaluation of the management of this case, the average scores given by Rater #1 were between 6.11 and 6.23, whereas those given by Rater #2 were between 5.23 and 5.58. Unlike other cases, there were statistically significant differences between the raters' correctness and safety scores in this case. This may be attributed to the higher number of questions and answers evaluated, the fact that the case was more complicated than the other cases, and the variability of clinician preferences in cases with more complex management. The questions and answers given in Figure 4 received the lowest score at all stages of the study. It was scored as such by raters because ChatGPT misjudged the serum Tg and Tg washout results.

The mean scores show that ChatGPT makes partially or largely correct, safe, and usable recommendations in the evaluation of cases. Despite this result, at least one answer in each case was evaluated by the raters as being partially or mostly incorrect, harmful, or not usable, indicating that the use of ChatGPT alone in case evaluations is inappropriate; however, its use as an additional assisting tool for physicians in clinical practice may be beneficial.

Technology is frequently used by patients and healthcare professionals to obtain information, and several previous studies have investigated the usability of these technology tools. In a 2015 review, it was determined that YouTube may provide false and misleading information to people seeking health-related information [9]. Other studies have evaluated the use of YouTube in obtaining information about thyroid diseases. A study examining YouTube content related to thyroid cancer concluded that the content was of low quality [10]. In addition, Dulak et al. found that the content related to hypothyroidism, which is a common ailment in the community, on YouTube is similarly of low quality and can lead to misdirection [11]. A study that examined the contents of radioactive iodine therapy stated that YouTube had both high-quality content that correctly informed patients, as well as content that contained unprofessional, incorrect, and incomplete information [12]. In a study evaluating thyroid cancer videos on TikTok, the quality of the videos was found to be unsatisfactory [13].

ChatGPT and similar AI applications are newer tools compared to YouTube and Google. As in every field, there have been many studies on the use of ChatGPT in medicine [2-4]. In a study where 22 case reports related to neuro-ophthalmic diseases were submitted to ChatGPT and the answers were evaluated, AI was found to help in making a fast and accurate diagnosis [14]. In another study conducted in 2023, ChatGPT was tested to predict drug-drug interactions, the results of which concluded that while ChatGPT is partially effective in this regard, it still requires improvement [15]. However, no previous study has investigated

whether ChatGPT can help patients or physicians with regard to thyroid diseases.

Our study showed that ChatGPT is an effective and reliable tool for patients to gain information. Compared to studies on other technology tools/platforms, such as YouTube, ChatGPT is a more appropriate source for patients to access correct and reliable information. In clinical practice, ChatGPT positions itself not as a practitioner but as an information provider. Although the information provided is mostly correct, reliable, and usable, it still needs to be checked by an expert in the relevant field, and the final decision regarding diagnosis or treatment should be made by the physician. The low-score responses in our study indicate that it is not yet appropriate to manage the patient using ChatGPT alone and that AI needs to be improved in this area.

One of the limitations of this study is that ChatGPT does not provide sources for its answers. Therefore, the use of the answers given may be unsuitable in the practice of evidence-based medicine. In addition, the fact that the study included information up to 2021 only means that the information may not be up-to-date. Another limitation is that ChatGPT can provide different answers to the same questions asked at different times.

Conclusions

Aside from the possibility that AI will play a significant role in shaping the future in various avenues of human life, the fact that it is so easily accessible, even today, necessitates further testing by professionals in various fields to evaluate its accuracy and safety. According to the findings of our study, ChatGPT can be used as an informative, useful, and safe resource for patients with thyroid nodules. Although ChatGPT is unsuitable as a primary resource for physicians, it has the potential to be a helpful and supportive tool if it is developed and kept up to date. However, more studies are required to validate our findings.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Ekin Y. Koroğlu, Sevgül Faki, Abbas A. Tam, Neslihan Çuhacı Seyrek, Oya Topaloglu, Reyhan Ersoy, Bekir Cakir

Acquisition, analysis, or interpretation of data: Ekin Y. Koroğlu, Nagihan Beştepe, Abbas A. Tam, Neslihan Çuhacı Seyrek

Drafting of the manuscript: Ekin Y. Koroğlu, Nagihan Beştepe

Critical review of the manuscript for important intellectual content: Ekin Y. Koroğlu, Sevgül Faki, Nagihan Beştepe, Abbas A. Tam, Neslihan Çuhacı Seyrek, Oya Topaloglu, Reyhan Ersoy, Bekir Cakir

Supervision: Abbas A. Tam, Oya Topaloglu, Reyhan Ersoy, Bekir Cakir

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue.

Animal subjects: All authors have confirmed that this study did not involve animal subjects or tissue.

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

- van Dis EA, Bollen J, Zuidema W, van Rooij R, Bockting CL: ChatGPT: five priorities for research. *Nature*. 2023, 614:224-6. [10.1038/d41586-023-00288-7](https://doi.org/10.1038/d41586-023-00288-7)
- Schulte B: Capacity of ChatGPT to identify guideline-based treatments for advanced solid tumors. *Cureus*. 2023, 15:e37938. [10.7759/cureus.37938](https://doi.org/10.7759/cureus.37938)
- Seth I, Sinkjær Kenney P, Bulloch G, Hunter-Smith DJ, Bo Thomsen J, Rozen WM: Artificial or augmented authorship? A conversation with a chatbot on base of thumb arthritis. *Plast Reconstr Surg Glob Open*. 2023, 11:e4999. [10.1097/GOX.0000000000004999](https://doi.org/10.1097/GOX.0000000000004999)
- Sharma S, Pajai S, Prasad R, Wanjari MB, Munjewar PK, Sharma R, Pathade A: A critical review of ChatGPT as a potential substitute for diabetes educators. *Cureus*. 2023, 15:e38380. [10.7759/cureus.38380](https://doi.org/10.7759/cureus.38380)
- Durante C, Grani G, Lamartina L, Filetti S, Mandel SJ, Cooper DS: The diagnosis and management of thyroid nodules: a review. *JAMA*. 2018, 319:914-24. [10.1001/jama.2018.0898](https://doi.org/10.1001/jama.2018.0898)

6. Grani G, Sponziello M, Pecce V, Ramundo V, Durante C: Contemporary thyroid nodule evaluation and management. *J Clin Endocrinol Metab.* 2020, 105:2869-83. [10.1210/clinem/dgaa322](https://doi.org/10.1210/clinem/dgaa322)
7. Haugen BR, Alexander EK, Bible KC, et al.: 2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: the American Thyroid Association guidelines task force on thyroid nodules and differentiated thyroid cancer. *Thyroid.* 2016, 26:1-133. [10.1089/thy.2015.0020](https://doi.org/10.1089/thy.2015.0020)
8. Baloch ZW, Asa SL, Barletta JA, et al.: Overview of the 2022 WHO classification of thyroid neoplasms. *Endocr Pathol.* 2022, 33:27-63. [10.1007/s12022-022-09707-3](https://doi.org/10.1007/s12022-022-09707-3)
9. Madathil KC, Rivera-Rodriguez AJ, Greenstein JS, Gramopadhye AK: Healthcare information on YouTube: a systematic review. *Health Informatics J.* 2015, 21:173-94. [10.1177/1460458213512220](https://doi.org/10.1177/1460458213512220)
10. Aydin MA, Akyol H: Quality of information available on YouTube videos pertaining to thyroid cancer. *J Cancer Educ.* 2020, 35:599-605. [10.1007/s13187-019-01502-9](https://doi.org/10.1007/s13187-019-01502-9)
11. Dułak NA, Anuszkiewicz K, Trzciński R, Fanciulli G, Stogowski P: YouTube as a patient-information source for hypothyroidism. *Minerva Endocrinol (Torino).* 2022, 1:3-7. [10.23736/S2724-6507.22.03732-0](https://doi.org/10.23736/S2724-6507.22.03732-0)
12. Şan H: Use of YouTube as an information source for radioactive iodine therapy: do YouTube videos have high quality?. *Mol Imaging Radionucl Ther.* 2022, 51:42-8. [10.4274/mirt.galenos.2021.85530](https://doi.org/10.4274/mirt.galenos.2021.85530)
13. Yang S, Zhan J, Xu X: Is TikTok a high-quality source of information on thyroid cancer?. *Endocrine.* 2023, 81:270-6. [10.1007/s12020-023-03332-8](https://doi.org/10.1007/s12020-023-03332-8)
14. Madadi Y, Delsoz M, Lao PA, Fong JW, Hollingsworth TJ, Kahook MY, Yousefi S: ChatGPT assisting diagnosis of neuro-ophthalmology diseases based on case reports. *medRxiv.* 2023, [10.1101/2023.09.13.23295508](https://doi.org/10.1101/2023.09.13.23295508)
15. Juhi A, Pipil N, Santra S, Mondal S, Behera JK, Mondal H: The capability of ChatGPT in predicting and explaining common drug-drug interactions. *Cureus.* 2023, 15:e36272. [10.7759/cureus.36272](https://doi.org/10.7759/cureus.36272)