

Article

The Role of Artificial Intelligence in Endocrine Management: Assessing ChatGPT's Responses to Prolactinoma Queries

Mustafa Can Şenoymak ^{1,*}, Nuriye Hale Erbatur ¹, İrem Şenoymak ² and Sevde Nur Fırat ³¹ Department of Endocrinology and Metabolism, University of Health Sciences Sultan, Abdulhamid Han Training and Research Hospital, Istanbul 34668, Turkey² Family Medicine Department, Üsküdar State Hospital, Istanbul 34662, Turkey³ Department of Endocrinology and Metabolism, University of Health Sciences, Ankara Training and Research Hospital, Ankara 06230, Turkey

* Correspondence: mustafacan.senoymak@sbu.edu.tr; Tel.: +90-535-317-89-59

Abstract: This research investigates the utility of Chat Generative Pre-trained Transformer (ChatGPT) in addressing patient inquiries related to hyperprolactinemia and prolactinoma. A set of 46 commonly asked questions from patients with prolactinoma were presented to ChatGPT and responses were evaluated for accuracy with a 6-point Likert scale (1: completely inaccurate to 6: completely accurate) and adequacy with a 5-point Likert scale (1: completely inadequate to 5: completely adequate). Two independent endocrinologists assessed the responses, based on international guidelines. Questions were categorized into groups including general information, diagnostic process, treatment process, follow-up, and pregnancy period. The median accuracy score was 6.0 (IQR, 5.4–6.0), and the adequacy score was 4.5 (IQR, 3.5–5.0). The lowest accuracy and adequacy score assigned by both evaluators was two. Significant agreement was observed between the evaluators, demonstrated by a weighted κ of 0.68 ($p = 0.08$) for accuracy and a κ of 0.66 ($p = 0.04$) for adequacy. The Kruskal–Wallis tests revealed statistically significant differences among the groups for accuracy ($p = 0.005$) and adequacy ($p = 0.023$). The pregnancy period group had the lowest accuracy score and both pregnancy period and follow-up groups had the lowest adequacy score. In conclusion, ChatGPT demonstrated commendable responses in addressing prolactinoma queries; however, certain limitations were observed, particularly in providing accurate information related to the pregnancy period, emphasizing the need for refining its capabilities in medical contexts.



Citation: Şenoymak, M.C.; Erbatur, N.H.; Şenoymak, İ.; Fırat, S.N. The Role of Artificial Intelligence in Endocrine Management: Assessing ChatGPT's Responses to Prolactinoma Queries. *J. Pers. Med.* **2024**, *14*, 330. <https://doi.org/10.3390/jpm14040330>

Academic Editor: Roberto Franceschi

Received: 21 February 2024

Revised: 15 March 2024

Accepted: 21 March 2024

Published: 22 March 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: artificial intelligence; ChatGPT; prolactinoma; hyperprolactinemia; health literacy

1. Introduction

Hyperprolactinemia is a clinical condition frequently encountered in endocrine practice, characterized by elevated serum prolactin levels, often posing challenges in terms of differential diagnosis and management [1]. While prolactinoma is a significant factor, the multitude of conditions leading to elevated prolactin levels, coupled with the requirement for a multidisciplinary approach due to prolactinoma's impact on various systems, gives rise to numerous questions and uncertainties among patients [2]. In these situations, patients commonly seek or are referred to endocrinology clinics, directing their inquiries to experts in the field. Responding to patients' questions about their conditions serves to alleviate unnecessary anxiety and expenses, while also streamlining disease management and improving treatment outcomes [3].

In the current context, artificial intelligence (AI) applications have become ubiquitous and easily accessible repositories of information. Artificial intelligence programs are widely used by both healthcare professionals and patients. As observed from these studies, it has been noted that patients find numerous advantages in the use of such applications. These advantages for patients include gaining a deeper understanding of their medical conditions through research conducted in artificial intelligence programs, as well as engaging in

interactive conversations by posing questions in a natural conversational manner, which they might not have had the opportunity to ask their doctors. Additionally, patients benefit from effectively managing their treatment through features like drug reminders, self-monitoring, dosage guidance, and accessing information on the potential side effects of therapy [4–6]. On the other hand, healthcare professionals benefit from utilizing artificial intelligence in the process of diagnostic radiology and medical education, as well as in the planning, execution, and composition of medical research within the field [7,8]. Despite the escalating frequency of its application, artificial intelligence in these domains is subject to notable limitations. The occasional lack of up-to-date information, variability in responses to different types of queries, and non-repeatability, resulting in varying responses at different times, can create misleading situations for patients. Additionally, it can be noted among its shortcomings that artificial intelligence in diagnostic processes may introduce bias, exhibit high error rates, and contribute to uniformity in research procedures [9,10].

The responses provided by an artificial intelligence program, Chat Generative Pre-trained Transformer (ChatGPT), to questions posed by patients in various fields have been extensively investigated in numerous studies [11–13]. Although results may vary, it has been observed in these studies that ChatGPT generally provides responses with high accuracy, regarding the relevant medical conditions. However, the use of artificial intelligence in prolactinoma has not been studied to date. Our study seeks to assess the adequacy and accuracy of responses generated by ChatGPT in addressing the most prevalent inquiries posed by patients with hyperprolactinemia and prolactinoma attending an endocrinology and metabolism department.

2. Materials and Methods

This research identified the 46 most frequently asked questions posed by patients seeking medical attention for hyperprolactinemia at an endocrinology and metabolism department. After compiling the list of questions, they were systematically presented to the artificial intelligence program ChatGPT (GPT3.5, version dated 13 January 2024) via <https://chat.openai.com> accessed on 13 January 2024. The recorded answers were then documented for subsequent analysis. In order to evaluate its reproducibility, each question was posed to ChatGPT twice.

The responses to inquiries were evaluated by two independent expert physicians working in the department of endocrinology and metabolism. The first reviewer is affiliated with the Department of Endocrinology and Metabolism at the University of Health Sciences, Ankara Training and Research Hospital and the second reviewer is affiliated with the Department of Endocrinology and Metabolism at the University of Health Sciences, İstanbul Sultan Abdülhamid Han Training and Research Hospital. Following this assessment, the responses were scored based on accuracy and adequacy, according to the international guidelines “Diagnosis and Treatment of Hyperprolactinemia: An Endocrine Society Clinical Practice Guideline” and “Guidelines of the Pituitary Society for the diagnosis and management of prolactinomas” [14–16]. The accuracy scale was operationalized as a 6-point Likert scale, with 1 denoting complete inaccuracy, 2 indicating a greater degree of inaccuracy than accuracy, 3 representing an approximate balance between accuracy and inaccuracy, 4 suggesting a higher accuracy level than inaccuracy, 5 reflecting near complete accuracy, and 6 signifying complete accuracy). Likewise, the adequacy scale was structured as a 5-point Likert scale, where 1 denoted complete inadequacy, 2 represented a greater inadequacy compared to adequacy, 3 signified an approximate equilibrium between adequacy and inadequacy, 4 indicated a higher degree of adequacy than inadequacy, and 5 indicated complete adequacy.

Statistical Analyses

Data analyses were conducted using IBM SPSS Statistics version 25.0 software. Outcome scores were presented descriptively, encompassing median [interquartile range (IQR)] values and mean [standard deviation (SD)] values. Group-wise comparisons were per-

formed using either the Mann–Whitney U test or the Kruskal–Wallis test (SPSS, version 25.0). Inter-rater concordance was evaluated utilizing the weighted κ statistic, covering a comprehensive range of scores, from 1 to 6 for accuracy and 1 to 5 for adequacy. Statistical significance was established at a threshold of $p < 0.05$. Significance values were adjusted by the Bonferroni correction for multiple tests. Responses to the repeated queries were subjected to comparison using the Wilcoxon signed rank test in order to assess reproducibility.

3. Results

Artificial intelligence was employed to address commonly encountered inquiries regarding prolactinoma posed by patients during routine endocrinology practice. Subsequently, two endocrinology and metabolism experts systematically assessed and assigned scores to the responses provided by ChatGPT (Table 1).

Table 1. Questions and average accuracy and adequacy scores given by medical doctors to the AI's responses ^a.

Questions	Accuracy Score ^b	Adequacy Score ^c
What is prolactin?	6	5
What is the physiological role of prolactin in the body?	6	5
How is prolactin measured?	6	3
What is the normal range for prolactin?	6	5
What precautions should be taken before blood sampling for prolactin measurement?	5.5	4
What factors contribute to elevated levels of prolactin?	6	5
Which medications can increase prolactin levels?	6	5
What are the symptoms of hyperprolactinemia?	6	4.5
What is macroprolactinemia?	6	4
What is prolactinoma?	6	5
What are the symptoms of prolactinoma?	6	4
How is the diagnosis of prolactinoma established?	5.5	3
In which gender is prolactinoma more commonly observed?	6	5
How can it be determined whether the elevated prolactin is due to a pituitary adenoma?	6	4
What is macroprolactinoma?	6	5
What is the treatment for prolactinoma?	6	5
Should every patient be treated, or is it possible to opt for a conservative approach with regular monitoring without intervention?	5	4.5
How is cabergoline used in the treatment of prolactinoma?	6	5
What are the side effects of cabergoline?	6	3
How is bromocriptine used in the treatment of prolactinoma?	5.5	4
What are the side effects of bromocriptine?	5	3
In what situations is surgery required for prolactinoma?	6	5
Is radiotherapy administered in the treatment of prolactinoma?	6	5
What should be the target prolactin level with treatment?	5	4.5
How long should medication be administered in the treatment of prolactinoma?	6	3
Do symptoms completely resolve after treatment?	5.5	3
Can prolactinoma shrink with medical treatment?	6	5
Is it possible to achieve complete recovery from prolactinoma after treatment?	6	4

Table 1. Cont.

Questions	Accuracy Score ^b	Adequacy Score ^c
How often should a patient diagnosed with prolactinoma consult a doctor?	5	2
How often should an MRI be performed in prolactinoma cases?	5	2.5
Can there be a recurrence after discontinuing medication for prolactinoma?	5.5	3.5
Can patients with prolactinoma conceive?	5	3
How is prolactinoma monitored during pregnancy?	2	3
How is prolactinoma treated during pregnancy?	2	3
In a pregnant patient with prolactinoma, when should medical treatment be discontinued?	5	3.5
Can women with prolactinoma use medication after childbirth?	5	3.5
Is breastfeeding allowed for a patient with prolactinoma after childbirth?	5	4
Should postmenopausal prolactinoma be treated?	5.5	4
What are the complications of elevated prolactin levels?	6	4.5
Which other hormones are affected by hyperprolactinemia?	5.5	5
What ocular manifestations are associated with prolactinoma?	6	5
Does prolactinoma cause headaches?	6	5
Does elevated prolactin affect sexual function in men?	6	5
Does hyperprolactinemia cause menstrual irregularities?	6	5
Is it possible to conceive while having elevated prolactin levels?	6	5

^a AI indicates artificial intelligence. ^b The accuracy scale was operationalized as a 6-point Likert scale, with 1 denoting complete inaccuracy, 2 indicating a greater degree of inaccuracy than accuracy, 3 representing an approximate balance between accuracy and inaccuracy, 4 suggesting a higher accuracy level than inaccuracy, 5 reflecting near complete accuracy, and 6 signifying complete accuracy. ^c The adequacy scale was structured as a 5-point Likert scale, where 1 denoted complete inadequacy, 2 represented a greater inadequacy compared to adequacy, 3 signified an approximate equilibrium between adequacy and inadequacy, 4 indicated a higher degree of adequacy than inadequacy, and 5 indicated complete adequacy.

Among the 46 questions under evaluation, the median average accuracy score demonstrated complete accuracy, registering at 6.0 (IQR, 5.4–6). The overall mean (SD) accuracy score of 5.5 (0.9) was positioned within the range spanning near complete accuracy and complete accuracy (Table 2).

Regarding adequacy, the average median score reached 4.5 (IQR, 3.5–5.0), indicating a level situated between a higher degree of adequacy than inadequacy and complete adequacy. The mean (SD) adequacy score of 4.2 (0.9) further substantiates this finding (Table 3).

According to the reproducibility test, the responses to the original and repeated questions did not differ significantly in terms of both accuracy and adequacy (with respective *p* values of 0.79 and 0.24, determined using the Wilcoxon signed rank test). The responses to repeated questions garnered a median accuracy score of 6 (IQR, 5.5–6.0; mean [SD] score, 5.5 [0.7]) and a median adequacy score of 4.5 (IQR, 4.0–5.0; mean [SD] score, 4.4 [0.7]). Despite variations in sentence structure and minimal changes, no major alterations in content were identified across the responses.

Evaluators demonstrated concordance, as evidenced by a weighted κ of 0.68 (*p* = 0.08) for accuracy and a substantial agreement reflected by a weighted κ of 0.66 (*p* = 0.04) for adequacy.

Table 2. Assessing accuracy ^a and comparing responses generated by AI through categorization into distinct question groups.

	General Information	Diagnostic Process	Treatment Process	Follow-Up	Pregnancy Period	Total	<i>p</i> Value
Rater 1							
Median (IQR)	6.0 (6.0–6.0)	6.0 (6.0–6.0)	6.0 (5.5–6.0)	6.0 (5.0–6.0)	5.0 (2.5–5.7)	6.0 (5.8–6.0)	0.001
Mean (SD)	6.0 (0.0)	6.0 (0.0)	5.8 (0.4)	5.7 (0.5)	4.4 (1.6)	5.6 (0.9)	
Rater 2							
Median (IQR)	6.0 (6.0–6.0)	6.0 (6.0–6.0)	6.0 (5.0–6.0)	5.0 (5.0–6.0)	5.0 (2.5–5.7)	6.0 (5.0–6.0)	0.015
Mean (SD)	5.8 (0.3)	5.8 (0.4)	5.5 (0.5)	5.4 (0.5)	4.4 (1.6)	5.5 (0.9)	
Average Score							
Median (IQR)	6.0 (6.0–6.0)	6.0 (6.0–6.0)	6.0 (5.2–6.0)	5.5 (5.0–6.0)	5.0 (2.5–5.7)	6.0 (5.4–6.0)	0.005
Mean (SD)	5.9 (0.2)	5.9 (0.2)	5.7 (0.4)	5.5 (0.5)	4.4 (1.6)	5.5 (0.9)	

^a The accuracy scale was operationalized as a 6-point Likert scale, with 1 denoting complete inaccuracy, 2 indicating a greater degree of inaccuracy than accuracy, 3 representing an approximate balance between accuracy and inaccuracy, 4 suggesting a higher accuracy level than inaccuracy, 5 reflecting near complete accuracy, and 6 signifying complete accuracy. Kruskal–Wallis (>2 variables) tests were used for non-parametric variables and data were given as median [interquartile range (IQR)] values and mean [standard deviation (SD)]. *p* values less than 0.05 were considered statistically significant. Significance values were adjusted using the Bonferroni correction for multiple tests.

Table 3. Scoring adequacy ^a and comparing responses generated by artificial intelligence based on categorization into question groups.

	General Information	Diagnostic Process	Treatment Process	Follow-Up	Pregnancy Period	Total	<i>p</i> Value
Rater 1							
Median (IQR)	5.0 (4.5–5.0)	5.0 (4.0–5.0)	5.0 (3.5–5.0)	4.0 (3.0–4.5)	3.0 (3.0–4.8)	5.0 (3.0–5.0)	0.018
Mean (SD)	4.8 (0.4)	4.6 (0.8)	4.4 (0.9)	3.6 (1.0)	3.6 (0.9)	4.3 (0.9)	
Rater2							
Median (IQR)	5.0 (4.0–5.0)	5.0 (4.0–5.0)	4.0 (3.5–5.0)	3.0 (2.5–4.5)	4.0 (3.0–4.8)	4.0 (3.0–5.0)	0.059
Mean (SD)	4.7 (0.5)	4.4 (0.8)	4.2 (0.8)	3.0 (1.1)	3.9 (0.8)	4.1 (0.9)	
Average Score							
Median (IQR)	5.0 (4.2–5.0)	5.0 (4.0–5.0)	4.5 (3.5–5.0)	3.5 (2.7–4.5)	3.5 (3.0–4.8)	4.5 (3.5–5.0)	0.023
Mean (SD)	4.9 (1.1)	4.5 (0.8)	4.3 (0.8)	3.5 (1.1)	3.8 (0.8)	4.2 (0.9)	

^a The adequacy scale was structured as a 5-point Likert scale, where 1 denoted complete inadequacy, 2 represented a greater inadequacy compared to adequacy, 3 signified an approximate equilibrium between adequacy and inadequacy, 4 indicated a higher degree of adequacy than inadequacy, and 5 indicated complete adequacy. Kruskal–Wallis (>2 variables) tests were used for non-parametric variables and data were given as median [interquartile range (IQR)] values and mean [standard deviation (SD)]. *p* values less than 0.05 were considered statistically significant. Significance values were adjusted using the Bonferroni correction for multiple tests.

Questions were stratified into distinct thematic categories, encompassing general information, diagnostic process, treatment process, follow-up, and the pregnancy period. The average median accuracy scores for these categories were 6.0 (IQR, 6.0–6.0), 6.0 (IQR,

6.0–6.0), 6.0 (IQR, 5.0–6.0), 5.0 (IQR, 5.0–6.0), and 5.0 (IQR, 2.5–5.7), respectively. The corresponding average mean [SD] scores were 5.9 [0.2], 5.9 [0.2], 5.7 [0.4], 5.5 [0.5], and 4.4 [1.6], respectively. Statistical analysis using the Kruskal–Wallis test ($p = 0.005$) indicated significant differences among these thematic groups. The subsequent pairwise post hoc Dunn test, employing Bonferroni adjustments, revealed significant distinctions, specifically for general information questions vs. pregnancy period questions ($p = 0.009$) and diagnostic process questions vs. pregnancy period questions ($p < 0.012$) (Table 2).

Evaluator #1 attributed the highest accuracy score (6.0) to 35 questions, constituting 76.1% of the total, while 8 questions (17.4%) garnered a rating of nearly completely accurate (2.0). Conversely, evaluator #2 bestowed the highest accuracy score (6.0) upon 28 questions (60.9%), with 15 questions (32.6%) characterized as nearly completely accurate (2.0). Both assessors assigned the lowest accuracy score (2.0) to the same two questions, signifying a notable disparity towards inaccuracy rather than accuracy (Figures 1–4).



Anonymous

4. What is the normal range for prolactin?



ChatGPT

The normal range for prolactin levels can vary slightly between different laboratories, and the units of measurement may also differ (nanograms per milliliter or micrograms per liter). Therefore, it's important to interpret results in the context of the specific laboratory's reference range. Generally, normal prolactin levels for non-pregnant adults are in the range of:

- For women: Approximately 2 to 29 ng/mL (or 2 to 29 µg/L)
- For men: Approximately 2 to 18 ng/mL (or 2 to 18 µg/L)

These values can provide a general guideline, but it's crucial to note that the normal range can be influenced by factors such as age, gender, menstrual cycle phase in women, and the specific laboratory's measurement techniques. Additionally, some individuals may have slightly higher or lower baseline prolactin levels without any underlying health issues.

If prolactin levels fall outside the normal range, it may indicate an underlying condition that needs further evaluation. High levels of prolactin (hyperprolactinemia) may be associated with conditions like pituitary tumors, hypothyroidism, or certain medications. Conversely, low levels of prolactin (hypoprolactinemia) are less common but may be seen in conditions affecting the hypothalamus or pituitary gland.

It's essential to discuss prolactin test results with a healthcare professional who can provide context based on the individual's medical history, symptoms, and other relevant factors. Abnormal prolactin levels may warrant additional investigations and, if necessary, appropriate management.

Figure 1. Samples of responses with the highest accuracy scores.

Evaluator #1 assessed the responses to 21 questions (45.7%) as entirely adequate, 13 (28.3%) as demonstrating a higher degree of adequacy than inadequacy, and 10 (21.7%) as approximately equal in terms of adequacy and inadequacy. For evaluator #2, the answers to 24 questions (52.2%) received a rating of complete adequacy, 10 (21.7%) were deemed

to have a higher degree of adequacy than inadequacy, and 11 (23.9%) were evaluated as approximately equal in terms of adequacy and inadequacy. Both evaluators assigned the lowest accuracy score of 2 and, notably, none of the questions received a score indicating complete inadequacy.

A compelling correlation between accuracy and adequacy was demonstrated, with a Spearman correlation coefficient (r) of 0.64 ($p < 0.001$), observed for all queried questions.

Similarly, when questions were categorized into distinct themes, including general information, diagnostic process, treatment process, follow-up, and the pregnancy period, the average median adequacy scores were 5.0 (IQR, 4.2–5.0), 5.0 (IQR, 4.0–5.0), 4.5 (IQR, 3.5–5.0), 3.5 (IQR, 2.7–4.5), and 3.5 (IQR, 3.0–4.8), respectively. The corresponding average mean [SD] scores were 4.9 [1.1], 4.5 [0.8], 4.3 [0.8], 3.5 [1.1], and 3.8 [0.8], respectively. While the Kruskal–Wallis test ($p = 0.023$) indicated statistically significant differences among the thematic groups, post hoc pairwise analysis did not reveal any significant distinctions (Table 3).



Anonymous

14. In which gender is prolactinoma more commonly observed?



ChatGPT

Prolactinomas can occur in both males and females, but they are more commonly observed in women. In fact, prolactinomas are one of the most common types of pituitary tumors in women of reproductive age.

The reasons for the higher prevalence of prolactinomas in women are not fully understood, but hormonal factors and the role of prolactin in reproductive function are likely contributors. Prolactinomas can lead to disruptions in the menstrual cycle, causing menstrual irregularities, amenorrhea (absence of menstruation), and fertility issues. The association between hormonal fluctuations, reproductive function, and the development of prolactinomas may contribute to the higher incidence in women.

However, it's important to note that prolactinomas can and do occur in men as well. Men with prolactinomas may experience symptoms such as sexual dysfunction, reduced libido, and, in some cases, galactorrhea (inappropriate milk production from the breasts). While prolactinomas are less common in men, they still represent a significant portion of pituitary tumors.

The diagnosis and management of prolactinomas are similar for both genders. Treatment options may include medications to lower prolactin levels (dopamine agonists like bromocriptine or cabergoline), surgery to remove the tumor (especially for larger tumors), or, in some cases, radiation therapy.

Regular monitoring and follow-up with healthcare professionals are essential to assess treatment effectiveness, manage symptoms, and monitor for potential complications. The prognosis for prolactinomas is generally favorable with appropriate management, and many individuals can achieve normal prolactin levels and resolution of symptoms.

Figure 2. Samples of responses with the highest accuracy scores.

**Anonymous**

34. How is prolactinoma monitored during pregnancy?

**ChatGPT**

Monitoring prolactinoma during pregnancy involves a combination of clinical assessments, hormonal monitoring, and, if necessary, imaging studies. The approach may vary based on individual factors, including the size and characteristics of the prolactinoma, the response to treatment, and the overall health of the pregnant individual. Here are key considerations for monitoring prolactinoma during pregnancy:

1. **Regular Follow-Up Appointments:**
 - Pregnant individuals with prolactinoma typically have regular follow-up appointments with their healthcare provider, often an endocrinologist or obstetrician. The frequency of these appointments may be determined based on individual needs and the specific characteristics of the prolactinoma.
2. **Hormonal Monitoring:**
 - Blood tests to monitor prolactin levels may be performed during pregnancy. Monitoring prolactin levels helps assess the effectiveness of treatment and ensures that levels remain within an acceptable range.
3. **Imaging Studies:**
 - In some cases, imaging studies such as magnetic resonance imaging (MRI) may be considered during pregnancy to assess the size and characteristics of the prolactinoma. The decision to perform imaging studies is typically based on specific clinical indications and considerations.
4. **Medication Adjustments:**
 - The management of prolactinoma during pregnancy may involve adjustments to medication dosages, if necessary. Dopamine agonist medications, such as cabergoline, are generally considered safe during pregnancy, but the decision to continue, modify, or discontinue medication is made on an individual basis.
5. **Assessment of Symptoms:**
 - Monitoring for symptoms related to prolactinoma, such as headaches, visual disturbances, or galactorrhea, is important during pregnancy. Any new or worsening symptoms should be promptly reported to healthcare providers.
6. **Collaborative Care:**
 - A collaborative approach involving endocrinologists, obstetricians, and other healthcare providers is crucial. Open communication ensures that the management plan is tailored to the specific needs and circumstances of the pregnant individual.

It's important to note that the approach to monitoring may vary among individuals, and decisions are made based on factors such as the individual's overall health, the size and characteristics of the prolactinoma, and the goals of treatment during pregnancy.

Pregnant individuals with prolactinoma should discuss their pregnancy plans and management with their healthcare provider early on. This allows for the development of a comprehensive care plan that addresses both the well-being of the pregnant individual and the management of the prolactinoma during this unique period. Regular follow-up and close communication with healthcare providers help ensure a smooth and well-managed

Figure 3. First answer with the lowest accuracy score.

**Anonymous**

35. How is prolactinoma treated during pregnancy?

**ChatGPT**

The treatment of prolactinoma during pregnancy requires careful consideration to balance the need for managing the prolactinoma with the safety of the developing fetus. Here are key considerations for the treatment of prolactinoma during pregnancy:

1. **Dopamine Agonist Medications:**
 - Dopamine agonist medications, such as bromocriptine and cabergoline, are the mainstay of treatment for prolactinoma. Among these, cabergoline is often preferred during pregnancy due to its longer duration of action and favorable safety profile.
2. **Safety of Cabergoline During Pregnancy:**
 - Cabergoline has been shown to be relatively safe during pregnancy when used at low doses to control prolactinoma. Studies have not demonstrated an increased risk of birth defects or adverse outcomes with the use of cabergoline during pregnancy. However, the decision to use any medication during pregnancy is made on an individual basis, weighing the potential benefits against potential risks.
3. **Monitoring Prolactin Levels:**
 - Regular monitoring of prolactin levels is typically performed during pregnancy to ensure that they remain within an acceptable range. This helps assess the effectiveness of treatment and guides any necessary adjustments to medication dosages.
4. **Imaging Studies:**
 - The decision to perform imaging studies, such as magnetic resonance imaging (MRI), during pregnancy depends on clinical indications. Imaging may be considered if there are specific concerns about the size or characteristics of the prolactinoma or if new symptoms arise.
5. **Individualized Care:**
 - Treatment plans are individualized based on factors such as the size and characteristics of the prolactinoma, the individual's overall health, and the goals of therapy during pregnancy. The healthcare provider will consider the potential risks and benefits of treatment in each specific case.
6. **Adjustments to Medication Dosages:**
 - Medication dosages may be adjusted based on the individual's response to treatment and changes in prolactin levels. Close monitoring allows for timely adjustments to optimize treatment during pregnancy.
7. **Consultation with a Maternal-Fetal Medicine Specialist:**
 - In complex cases or when there are concerns about the impact of prolactinoma or its treatment on the pregnancy, consultation with a maternal-fetal medicine specialist may be considered. This specialist can provide additional expertise in managing high-risk pregnancies.

It's important for pregnant individuals with prolactinoma to communicate their pregnancy plans with their healthcare provider early on. The decision-making process involves a collaborative approach, and the healthcare provider will work closely with the individual to develop a comprehensive care plan that addresses both the management of the prolactinoma and the well-being of the developing fetus.

Regular follow-up appointments, close monitoring, and open communication with healthcare providers are crucial to ensure the best possible outcomes during pregnancy for both the mother and the baby.

Figure 4. Second answer with the lowest accuracy score.

4. Discussion

This study marks a pioneering exploration into the utilization of an artificial intelligence tool, ChatGPT, for addressing inquiries related to hyperprolactinemia and prolactinoma. To the best of our knowledge, it serves as the first comprehensive study systematically examining responses generated by ChatGPT, specifically within the realm of prolactinoma. The evaluation of the effectiveness of artificial intelligence in endocrinological patient communication through the two crucial dimensions of accuracy and adequacy revealed noteworthy findings.

In contemporary medical practices, a crucial aspect of successful disease management involves patients being aware of their conditions and actively participating in the treatment and management process. Studies have indicated a positive association between increased health literacy and improved disease outcomes [17,18]. Therefore, various approaches are being explored to enhance health literacy. Technology, particularly internet-based tools, is increasingly replacing traditional educational and informational resources to promote health literacy [19]. The emergence of artificial intelligence, notably as the most accessible and popular among these technologies, underscores its widespread use in various fields. In our study, we utilized ChatGPT as an artificial intelligence tool, which is the most frequently used and popular, and we investigated its role in the management of prolactinoma.

Prolactinoma is a commonly encountered condition in endocrinology practice, characterized by several pitfalls in diagnosis and management. The rationale behind selecting prolactinoma for this study stems from the need for patients to obtain information in this confusing disease, which necessitates a thorough evaluation and is commonly accompanied by a multitude of inquiries. Hence, there is a necessity for a supportive tool to assist patients in acquiring knowledge. In our study, the questions were compiled by actively practicing endocrinologists and consisted of inquiries commonly posed by patients in the outpatient clinic setting. The content and techniques of the questions were structured to mimic those typically encountered in clinical practice (Table 1). This approach enhances the alignment of our study with real-life data, ultimately bridging real-world data with an artificial intelligence platform.

In our study, it was found that ChatGPT demonstrated a commendable performance in offering accurate and adequate responses to the questions posed by patients seeking information on hyperprolactinemia and prolactinoma. Upon subjecting it to group analysis, specific areas of concern became apparent and, notably, a lower accuracy in information related to pregnancy was revealed. According to the Endocrine Society Clinical practice guideline and guidelines of the Pituitary Society for the diagnosis and management of prolactinomas, in prolactinoma cases during pregnancy, it is generally recommended to discontinue dopamine agonist therapy and ensure close monitoring, aimed at minimizing fetal exposure to the medication. However, exceptions exist for cases presenting with symptoms indicative of mass effect or those involving macroprolactinomas. The routine monitoring of prolactin levels is discouraged during pregnancy, due to the inherent ambiguity in interpreting such levels in pregnant individuals. Instead, emphasis is placed on regular clinical evaluations and the assessment of visual fields as part of the follow-up regimen for optimal management [14–16]. Despite these established guidelines, our study highlights that ChatGPT's responses on this topic, particularly those suggesting prolactin measurement and continuing medical treatment during pregnancy, can be misleading for patients (Figures 3 and 4). The observed misguidance in ChatGPT's responses underscores the importance of the cautious interpretation of AI-generated information in medical contexts. It emphasizes the need for the ongoing refinement and validation of AI models to align more closely with evolving clinical knowledge and guidelines. Additionally, this finding serves as a reminder of the critical role that human expertise plays in interpreting AI-generated information, especially in complex medical scenarios where context and individual patient factors are paramount.

While there may be a limited body of research specifically addressing prolactinoma, the effectiveness of artificial intelligence in addressing patient queries has been investigated

in numerous studies spanning various diseases and medical conditions [11–13,20,21]. In a study conducted by Goodman and colleagues, 33 physicians from different medical specialties posed 284 questions to ChatGPT and the responses were evaluated for accuracy and completeness, similar to our study. The accuracy of the responses was assessed between almost completely and completely correct, and completeness was evaluated as being complete and comprehensive [20]. In another study related to bariatric surgery, ChatGPT was found to answer 86.8% of the questions; however, unlike our study, proficiency and accuracy were not separately evaluated and a common scoring system was employed [22].

Studies on the success of artificial intelligence programs yield conflicting results. While many studies report successful outcomes, some have resulted in failure. For instance, in a study conducted by Rahsepar and colleagues, it was noted that ChatGPT provided correct answers to 70% of 120 questions related to lung cancer but faced criticism for not achieving complete accuracy. Similarly, questions pertaining to hepatocellular carcinoma and cirrhosis were answered with over 70% accuracy, but the comprehensiveness remained at around 40%, indicating a need for improvement [12]. Recent research has compared the performance of ChatGPT's old and new versions in responding to medical queries, revealing that the use of newer versions has increased the success rate [23].

A primary constraint within our investigation, as well as a significant quandary concerning artificial intelligence systems, pertains to the variability in responses contingent upon question formulation. This variability, induced by discrepancies in patients' articulatory skills when formulating inquiries, could yield divergent responses. Additionally, despite the appraisal of responses by healthcare experts, uncertainties persist regarding whether these responses will exert an analogous impact on the patients. Consequently, patients may not fully avail themselves of precise and comprehensive information, as observed within such investigations. To alleviate this challenge, multicenter studies should be conducted, involving patients with diverse sociocultural backgrounds, educational proficiencies, and linguistic diversities, wherein the data evaluated by healthcare professionals should also be concurrently assessed by patients.

Some other constraints impact the robustness of our study. Firstly, ChatGPT's reliance on information up to 2022 raises concerns about its currency and the potential incorporation of outdated knowledge. Secondly, the subjective nature of adequacy evaluations introduces potential bias, as responses were not objectively measured in terms of appropriateness. Additionally, the evaluation solely focused on ChatGPT, overlooking the various other artificial intelligence platforms available for use in the medical field. Comparative evaluations of these platforms could significantly enhance the study, enabling patients to find the most appropriate platform for their healthcare needs.

5. Conclusions

This study underscores the commendable performance of ChatGPT in addressing patient inquiries related to hyperprolactinemia and prolactinoma and highlights the positive potential of artificial intelligence in enhancing patient communication and promoting health literacy. However, the findings also reveal specific areas of concern, particularly in the realm of pregnancy-related information, where ChatGPT exhibited a lower accuracy. This highlights the critical need for the careful interpretation of artificial intelligence-generated information in intricate medical situations, emphasizing the importance of the continuous refinement and validation of AI models, as well as the pivotal role of human expertise.

Author Contributions: Conceptualization, İ.Ş. and S.N.F.; Methodology, M.C.Ş. and İ.Ş.; Formal Analysis, N.H.E.; Investigation, N.H.E. and M.C.Ş.; Data Curation, M.C.Ş. and N.H.E.; Writing—Original Draft Preparation, İ.Ş., M.C.Ş. and S.N.F.; Writing—Review and Editing, S.N.F. and M.C.Ş.; Visualization, N.H.E.; Supervision, S.N.F.; Project Administration, İ.Ş. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study, as it does not entail the utilization of human subjects or patient data.

Data Availability Statement: The data supporting the findings of this study are available upon reasonable request to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Vilar, L.; Vilar, C.F.; Lyra, R.; da Freitas, M.C. Pitfalls in the Diagnostic Evaluation of Hyperprolactinemia. *Neuroendocrinology* **2019**, *109*, 7–19. [\[CrossRef\]](#)
2. Auriemma, R.S.; Pirchio, R.; Pivonello, C.; Garifalos, F.; Colao, A.; Pivonello, R. Approach to the Patient with Prolactinoma. *J. Clin. Endocrinol. Metab.* **2023**, *108*, 2400–2423. [\[CrossRef\]](#)
3. Riedl, D.; Schüssler, G. The Influence of Doctor-Patient Communication on Health Outcomes: A Systematic Review. *Z. Psychosom. Med. Psychother.* **2017**, *63*, 131–150. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Ramezani, M.; Takian, A.; Bakhtiari, A.; Rabiee, H.R.; Ghazanfari, S.; Mostafavi, H. The Application of Artificial Intelligence in Health Policy: A Scoping Review. *BMC Health Serv. Res.* **2023**, *23*, 1416. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Dave, T.; Athaluri, S.A.; Singh, S. ChatGPT in Medicine: An Overview of Its Applications, Advantages, Limitations, Future Prospects, and Ethical Considerations. *Front. Artif. Intell.* **2023**, *6*, 1169595. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Khodve, G.B.; Banerjee, S. Artificial Intelligence in Efficient Diabetes Care. *Curr. Diabetes Rev.* **2023**, *19*, e050922208561. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Mese, I.; Taslicay, C.A.; Sivrioglu, A.K. Improving radiology workflow using ChatGPT and artificial intelligence. *Clin. Imaging* **2023**, *103*, 109993. [\[CrossRef\]](#)
8. Ruksakulpiwat, S.; Kumar, A.; Ajibade, A. Using ChatGPT in Medical Research: Current Status and Future Directions. *J. Multidiscip. Healthc.* **2023**, *16*, 1513–1520. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Abbasgholizadeh Rahimi, S.; Légaré, F.; Sharma, G.; Archambault, P.; Zomahoun, H.T.V.; Chandavong, S.; Rheault, N.; Wong, T.; Langlois, L.; Couturier, Y.; et al. Application of Artificial Intelligence in Community-Based Primary Health Care: Systematic Scoping Review and Critical Appraisal. *J. Med. Internet Res.* **2021**, *23*, e29839. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Mohammad, B.; Supti, T.; Alzubaidi, M.; Shah, H.; Alam, T.; Shah, Z.; Househ, M. The Pros and Cons of Using ChatGPT in Medical Education: A Scoping Review. *Stud. Health Technol. Inform.* **2023**, *305*, 644–647. [\[CrossRef\]](#)
11. Mondal, H.; Dash, I.; Mondal, S.; Behera, J.K. ChatGPT in Answering Queries Related to Lifestyle-Related Diseases and Disorders. *Cureus* **2023**, *15*, e48296. [\[CrossRef\]](#)
12. Yeo, Y.H.; Samaan, J.S.; Ng, W.H.; Ting, P.-S.; Trivedi, H.; Vipani, A.; Ayoub, W.; Yang, J.D.; Liran, O.; Spiegel, B.; et al. Assessing the Performance of ChatGPT in Answering Questions Regarding Cirrhosis and Hepatocellular Carcinoma. *Clin. Mol. Hepatol.* **2023**, *29*, 721–732. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Rahsepar, A.A.; Tavakoli, N.; Kim, G.H.J.; Hassani, C.; Abtin, F.; Bedayat, A. How AI Responds to Common Lung Cancer Questions: ChatGPT versus Google Bard. *Radiology* **2023**, *307*, e230922. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Melmed, S.; Casanueva, F.F.; Hoffman, A.R.; Kleinberg, D.L.; Montori, V.M.; Schlechte, J.A.; Wass, J.A.H. Diagnosis and Treatment of Hyperprolactinemia: An Endocrine Society Clinical Practice Guideline. *J. Clin. Endocrinol. Metab.* **2011**, *96*, 273–288. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Casanueva, F.F.; Molitch, M.E.; Schlechte, J.A.; Abs, R.; Bonert, V.; Bronstein, M.D.; Brue, T.; Cappabianca, P.; Colao, A.; Fahlbusch, R.; et al. Guidelines of the Pituitary Society for the Diagnosis and Management of Prolactinomas. *Clin. Endocrinol.* **2006**, *65*, 265–273. [\[CrossRef\]](#)
16. Petersenn, S.; Fleseriu, M.; Casanueva, F.F.; Giustina, A.; Biermasz, N.; Biller, B.M.K.; Bronstein, M.; Chanson, P.; Fukuoka, H.; Gadelha, M.; et al. Diagnosis and Management of Prolactin-Secreting Pituitary Adenomas: A Pituitary Society International Consensus Statement. *Nat. Rev. Endocrinol.* **2023**, *19*, 722–740. [\[CrossRef\]](#)
17. Sierra, M.; Cianelli, R. Health Literacy in Relation to Health Outcomes: A Concept Analysis. *Nurs. Sci. Q.* **2019**, *32*, 299–305. [\[CrossRef\]](#)
18. Kim, M.T.; Kim, K.B.; Ko, J.; Murry, N.; Xie, B.; Radhakrishnan, K.; Han, H.-R. Health Literacy and Outcomes of a Community-Based Self-Help Intervention: A Case of Korean Americans with Type 2 Diabetes. *Nurs. Res.* **2020**, *69*, 210–218. [\[CrossRef\]](#)
19. Kim, K.; Shin, S.; Kim, S.; Lee, E. The Relation Between eHealth Literacy and Health-Related Behaviors: Systematic Review and Meta-Analysis. *J. Med. Internet Res.* **2023**, *25*, e40778. [\[CrossRef\]](#)
20. Goodman, R.S.; Patrinely, J.R.; Stone, C.A.; Zimmerman, E.; Donald, R.R.; Chang, S.S.; Berkowitz, S.T.; Finn, A.P.; Jahangir, E.; Scoville, E.A.; et al. Accuracy and Reliability of Chatbot Responses to Physician Questions. *JAMA Netw. Open* **2023**, *6*, e2336483. [\[CrossRef\]](#)
21. Ali, R.; Tang, O.Y.; Connolly, I.D.; Fridley, J.S.; Shin, J.H.; Zadnik Sullivan, P.L.; Cielo, D.; Oyelese, A.A.; Doberstein, C.E.; Telfeian, A.E.; et al. Performance of ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards Preparation Question Bank. *Neurosurgery* **2023**, *93*, 1090–1098. [\[CrossRef\]](#) [\[PubMed\]](#)

22. Samaan, J.S.; Yeo, Y.H.; Rajeev, N.; Hawley, L.; Abel, S.; Ng, W.H.; Srinivasan, N.; Park, J.; Burch, M.; Watson, R.; et al. Assessing the Accuracy of Responses by the Language Model ChatGPT to Questions Regarding Bariatric Surgery. *Obes. Surg.* **2023**, *33*, 1790–1796. [[CrossRef](#)] [[PubMed](#)]
23. Takagi, S.; Watari, T.; Erabi, A.; Sakaguchi, K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: Comparison Study. *JMIR Med. Educ.* **2023**, *9*, e48002. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.