EARLY RELEASE / **ORIGINAL ARTICLE**

Infect Dis Clin Microbiol 2024; 6(1)

How Reliable is ChatGPT as a Novel Consultant in Infectious Diseases and Clinical Microbiology?

Gülşah Tunçer[1], Kadir Görkem Güçlü[2]

[1] Bilecik Training and Research Hospital, Bilecik, Türkiye,
[2] İstanbul Haseki Training and Research Hospital, İstanbul, Türkiye

ABSTRACT

Objective: The study aimed to investigate the reliability of ChatGPT's answers to medical questions, including those sourced from patients and guide recommendations. The focus was on evaluating ChatGPT's accuracy in responding to various types of infectious disease questions.

Materials and Methods: The study was conducted using 200 questions sourced from social media, experts, and guidelines related to various infectious diseases like urinary tract infection, pneumonia, HIV, various types of hepatitis, COVID-19, skin infections, and tuberculosis. The questions were arranged for clarity and consistency by excluding repetitive or unclear ones. The answers were based on guidelines from reputable sources like the Infectious Diseases Society of America (IDSA), Centers for Disease Control and Prevention (CDC), European Association for the Study of Liver Disease (EASL) and Joint United Nations Programme on HIV/AIDS (UNAIDS) AIDSinfo. According to the scoring system, completely correct answers were given 1-point, and completely incorrect ones were given 4-points. To assess reproducibility, each question was posed twice on separate computers. Repeatability was determined by the consistency of the answers' scores.

Results: In the study, ChatGPT was posed with 200 questions: 107 from social media platforms and 93 from guidelines. The questions covered a range of topics: urinary tract infections (n=18 questions), pneumonia (n=22), HIV (n=39), hepatitis B and C (n=53), COVID-19 (n=11), skin and soft tissue infections (n=38), and tuberculosis (n=19). The lowest accuracy was 72% for urinary tract infections. ChatGPT answered 92% of social media platform questions correctly (scored 1-point) versus 69% of guideline questions (p=0.001; OR=5.48, 95% CI=2.29-13.11).

Conclusion: Artificial intelligence is widely used in the medical field by both healthcare professionals and patients. Although ChatGPT answers questions from social media platforms quite properly, we recommend that healthcare professionals be conscientious when using it.

Keywords: infectious disease, Chat-GTP, medical questions

Corresponding Author:
Kadir Görkem Güçlü

E-mail:
gorkemguclurd@gmail.com

Received: October 24, 2023
Accepted: December 14, 2023
Published: February 16, 2024

Suggested citation:
Tunçer G, Güçlü KG. How Reliable is ChatGPT as a Novel Consultant in Infectious Diseases and Clinical Microbiology? Infect Dis Clin Microbiol. 2024;1.

DOI: 10.36519/idcm.2024.286

INTRODUCTION

Artificial intelligence models have influenced many branches of science in recent years. It is used in various departments of medicine. ChatGPT (Chat Generative Pre-trained Transformer) is a text-based artificial intelligence model developed by OpenAI (1). ChatGPT can be used in many areas of medicine, such as to generate medical text, answer medical questions, provide recommendations for diagnosis and treatment, translate medical documents, and have a medical conversation (2).

The use of artificial intelligence in the medical field is increasing day by day. Patients often turn to the internet and social media platforms for quick answers to their medical concerns. Nevertheless, evaluating the quality of the information these platforms provide is very important. Unlike social media platforms, ChatGPT is a system that blends information by accessing it from many reliable sources. Despite having limited access to medical data, ChatGPT performs at the level of a third-year medical student in licensing exams, encouraging discussions on emergency medicine within medicine (3). For example, pediatric urology questions were answered very well in a study conducted using text-based artificial intelligence modeling (5). Although ChatGPT is thought to be promising in producing consistent responses, it is important to determine the accuracy of the medical information it provides. Artificial intelligence can cause many misdirections that cause an information epidemic called "infodemic," which can also threaten public health (4).

There are few studies on the use of artificial intelligence in the field of medicine. To our knowledge, our study is the first on this subject in the field of infectious diseases in Türkiye. We aimed to investigate the reliability and accuracy of ChatGPT's answers to questions about infectious diseases.

MATERIALS AND METHODS

A total of 200 questions were collected from social media platforms (YouTube, X [formerly named Twitter], Facebook), questions directed to infectious disease societies and specialists, or infectious disease guidelines. A social media question was defined as a question derived from social media. A guideline question was defined as a question derived from various infectious diseases guidelines. Of those, 93 were obtained from social media platforms and 107 from guidelines. The questions and the responses given by the experts are shown in the Supplementary Table. Questions about urinary tract infection, pneumonia, HIV, hepatitis B, hepatitis C, COVID-19, skin and soft tissue infections, and tuberculosis were included in the study. Our study did not include patient data, so ethics committee approval was not received.

Social media platform questions were selected from the questions posed to infectious disease associations and experts via social media platforms between 1 and 30 September 2023. Questions that were repetitive, with grammatical errors and unclear answers were not included. The responses to the guideline questions were obtained from the Infectious Diseases Society of America (IDSA), Centers for Disease Control and Prevention (CDC), European Association for the Study of Liver Disease (EASL) and Joint United Nations Programme on HIV/AIDS (UNAIDS) AIDSinfo guidelines, Turkish Thoracic Society Community-Acquired Pneumonia and Tuberculosis guidelines and Skin-Soft Tissue Infections Consensus Report. Responses to questions prepared from the guidelines were mostly 'high level of evidence' and 'strong recommendation'. In addition, questions covering the main topics were asked using the questions prepared from international and national guidelines. The questions included in the study were prepared in Turkish. They were asked ChatGPT in Turkish, and answers were obtained in Turkish from ChatGPT. The questions were translated into English to be included in the Supplementary Table in this manuscript.

HIGHLIGHTS

- While 92% of the questions from social media platforms received a 1-point response, 69% from guidelines received a 1-point response.
- ChatGPT achieved the highest correct answer rate in tuberculosis questions.
- When 1 and 2-points answers are evaluated together; ChatGPT answered social media platform questions more accurately than guide questions.

Table 1. Scoring the responses to guideline and social media platform questions.

Parameters	Guideline questions	Social media platforms questions	Score				
			1-point	2-points	3-points	4-points	Total
Total (n, %)	107 (53.5)	93 (46.5)	160 (80)	28 (14)	8 (4)	4 (2)	200
Urinary tract infections (n, %)	10 (55.5)	8 (44.5)	13 (72.3)	4 (22.2)	1 (5.5)	0 (0)	18
Pneumoniae (n, %)	11 (50)	11 (50)	17 (77.3)	2 (9)	3 (13.7)	0 (0)	22
HIV (n, %)	17 (43.6)	22 (56.4)	29 (74.4)	8 (20.5)	2 (5.1)	0 (0)	39
Hepatitis B (n, %)	25 (69.4)	11 (30.6)	30 (83.3)	5 (13.9)	0 (0)	1 (2.8)	36
Hepatitis C (n, %)	12 (70.6)	5 (29.4)	13 (76.5)	3 (17.6)	0 (0)	1 (5.9)	17
COVID-19 (n, %)	1 (9)	10 (91)	9 (82)	2 (18)	0 (0)	0 (0)	11
Skin-soft tissue infections (n, %)	31 (81.6)	7 (18.4)	32 (84.2)	3 (7.9)	2 (5.3)	1 (2.6)	38
Tuberculosis (n, %)	0 (0)	19 (100)	17 (90)	1 (5)	0 (0)	1 (5)	19
Guideline	107 (100)	0 (0)	74 (69.2)	23 (21.5)	8 (7.5)	2 (1.8)	107
Social Media	0 (0)	93 (100)	86 (92.5)	5 (5.4)	0 (0)	2 (2.1)	93

The completely correct answers to the questions were evaluated as 1-point, correct but insufficient answers 2-points, mixed or misleading answers 3-points, and completely wrong answers 4-points. To evaluate the reproducibility of the answers, each question was asked twice on different computers. Repeatability was defined as the consistency of two answers with similar rating categorizations. If the answer given to the repeated question was in a different score category or contained information at a different level of detail, it was considered negative in terms of repeatability. The answers were evaluated by two separate infectious disease specialists on two separate computers with the ChatGPT-4-September Update version. The final decision was made after the different answers were evaluated by a third expert. Those who got 1 and 2-points from the social media platforms and guideline questions were compared with the chi-square test to assess whether or not there was a significant difference between the answers.

We conducted statistical analyses using MS Excel 16.0 (Microsoft Corp., USA). The scores assigned to the answers were presented as percentages. Categorical data were presented as numbers and percentages. The chi-square test was used to compare categorical data. The statistical significance was set as $p<0.05$.

RESULTS

In our study, a total of 200 questions were asked to ChatGPT, of which 107 were from social media platforms, and 93 were from guidelines. Eighteen questions were asked about urinary tract infection, 22 about pneumonia, 39 about HIV, 53 about hepatitis B and C, 11 about COVID-19, 38 about skin and soft tissue infection, and 19 about tuberculosis (Table 1).

The highest correct answer rate of 90% was achieved in questions on tuberculosis, as 17 of 19 answers got 1-point. Regarding urinary tract infections, ChatGPT provided the least accurate response (72%). The mean±standard deviation (SD) score for ChatGPT-generated responses was 1.11±0.48 for the point of questions from social media platforms. Guidelines questions point mean±SD score for ChatGPT-generated responses was 1.42±0.71. While 92% of the questions from social media platforms received a 1-point response, 69% from guidelines received a 1-point response. When 1-point answers between both question groups were compared, the difference was statistically significant ($p=0.001$; OR=5.48, 95% CI=2.29-13.11) (Table 1).

A total of 86 (92.5%) of the answers given to 93 social media platform questions were evaluated as

Table 2. Comparison of the rates of correct responses for guideline and social media platform questions.

Parameters	Total	Guideline (n=107)	Social media (n=93)	OR	95% CI	*p*
1-point (n, %)	160 (100)	74 (46.3)	86 (53.7)	5.48	2.29-13.11	0.001
1 and 2-points (n, %)	188 (100)	97 (90.7)	91 (97.8)	4.69	1.00-21.98	0.049

OR: Odd's ratio, **CI:** Confidence interval.

1-point, and 5 (5.4%) were evaluated as 2-points. Of the 107 guideline questions, 74 (69.2%) received 1-point, and 23 (21.5%) received 2-points. When the answers with 1 and 2-points were evaluated together, the difference between the social media platform questions and the guideline questions was significant (97.8% vs 90.7%; *p*=0.049; OR=4.69, 95% CI=1.00-21.98) (Table 2).

DISCUSSION

ChatGPT has a wide information network and generally answers medical questions accurately. Our study investigated the reliability of ChatGpT's answers to questions, including patients' questions and guide recommendations. ChatGPT provides correct answers to questions about different types of infections at different rates. The high accuracy rate, especially for tuberculosis, shows that ChatGPT provides more accurate information about certain types of infections; however, it gave poorer answers to questions about urinary infections. Although there were differences in accuracy rates between subjects, in our study, the rates of questions receiving 1 and 2-points in both question groups were over 90%. This showed that ChatGPT's rate of correct answers was high, although there were deficiencies in some answers.

A similar rate was obtained in the study of Çaglar et al., including 137 questions about pediatric urology; 92% of the questions were answered correctly by ChatGPT (5). The same study stated that 5.1% of the responses to all questions were correct but insufficient, and 2.9% contained correct information along with misleading information (5). In another study conducted in South Korea, in which 79 medical school exam questions were evaluated using the ChatGPT January 1, 2023 version, it was observed that the performance of ChatGPT was lower than that of medical students (6). In another study conducted in 2023, it was observed that ChatGPT improved clinical workflow and radiology services in radiological decision-making (7). Dave et al. demonstrated the feasibility of using this potential of artificial intelligence in the field of medicine (8).

Nevertheless, some studies in the literature show that ChatGPT cannot provide appropriate answers to medical questions. It has been shown that the accuracy level of ChatGPT may be affected by the quality of information sources, with a significant difference between social media platforms and guideline questions. In this regard, it can be said that text-based artificial intelligence should obtain information from more reliable sources. Although ChatGPT answered social media platform questions at higher rates than guideline questions, its performance may have decreased because the directory information was more complex and specific.

After all, the rate of correct answers of ChatGPT to the guideline questions was relatively high, but healthcare professionals should be careful while using it. The accuracy of the information provided by ChatGPT must be checked with the guidelines. A recent study by Singhal et al. indicated that programs specific to the medical field have been developing (9). They reported that using a combination of prompting strategies, Flan-PaLM achieves state-of-the-art accuracy on every MultiMedQA multiple-choice dataset (MedQA3, MedMCQA4, PubMedQA5 and Measuring Massive Multitask Language Understanding [MMLU] clinical topics6), including 67.6% accuracy on MedQA (US Medical Licensing Exam-style questions), surpassing the prior state of the art by more than 17%. Nonetheless, they concluded the resulting model, Med-PaLM, performs promisingly but remains inferior to clinicians (9).

In another study, ChatGPT interpreted the clinical evaluation of 36 patients, and an overall accuracy of 71.7% was achieved (10). While ChatGPT showed high performance with 76.9% accuracy in making the final diagnosis, it showed the lowest performance with 60% accuracy in creating a differential diagnosis. ChatGPT showed poorer performance on differential diagnosis-type questions than answering questions about general medical information (p=0.02) (10). It may be concluded that ChatGPT lacks the medical expertise and context required to fully understand the complex relationship between different conditions and treatments (11). Although it is a powerful text-based artificial intelligence model, it has some limitations, such as reasoning, establishing context, and limited text input. The inability to establish as comprehensive a context as humans and the reliability of sources may not always produce correct answers.

In conclusion, healthcare professionals and patients widely use artificial intelligence in the medical field. Although ChatGPT answers social media questions well, we recommend that healthcare professionals be conscientious when using ChatGPT. Given these considerations, the direct use of ChatGPT in the field of infectious diseases carries associated risks in its current state and necessitates active verification of information by users. Although there were certain limitations specific to infectious disease medicine, the results of this study indicated that ChatGPT's medical knowledge has expanded and implies its potential to handle specific medical questions in the future.

Ethical Approval: N.A.

Informed Consent: N.A.

Peer-review: Externally peer-reviewed

Author Contributions: Concept – G.T., K.G.G.; Design – G.T., K.G.G.; Supervision – G.T., K.G.G.; Fundings – G.T., K.G.G.; Materials – G.T., K.G.G.; Data Collection and/or Processing – G.T., K.G.G.; Analysis and/or Interpretation – G.T., K.G.G.; Literature Review – G.T., K.G.G.; Writer – G.T., K.G.G.; Critical Reviews – G.T., K.G.G.

Conflict of Interest: The authors declare no conflict of interest.

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

Acknowledgement: We thank Dr. Serkan Sürme who is a specialist in infectious diseases and clinical microbiology for his contributions.

REFERENCES

1 Introducing ChatGPT, November 30, 2022 [Internet]. OpenAI. [cited October 24, 2023). Available from: https://openai.com/blog/chatgpt/

2 Peksen A, ChatGPT. Using ChatGPT in the medical field: a narrative Infect Dis Clin Microbiol. 2023;1:66-8. [CrossRef]

3 Howard A, Hope W, Gerada A. ChatGPT and antimicrobial advice: the end of the consulting infection doctor? Lancet Infect Dis. 2023;23(4):405-6. [CrossRef]

4 De Angelis L, Baglivo F, Arzilli G, Privitera GP, Ferragina P, Tozzi AE, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front Public Health. 2023;11:1166120. [CrossRef]

5 Caglar U, Yildiz O, Meric A, Ayranci A, Gelmis M, Sarilar O, Ozgor F. Evaluating the performance of ChatGPT in answering questions related to pediatric urology. J Pediatr Urol. 2023:S1477-5131(23)00318-2. [CrossRef]

6 Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. J Educ Eval Health Prof. 2023;20:1. [CrossRef]

7 Rao A, Kim J, Kamineni M, Pang M, Lie W, Succi MD. Evaluating ChatGPT as an adjunct for radiologic decision-making. medRxiv [Preprint]. 2023 Feb 7:2023.02.02.23285399. Update in: J Am Coll Radiol. 2023 Jun 21. [CrossRef]

8 Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023;6:1169595. [CrossRef]

9 Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-80. Erratum in: Nature. 2023 Jul 27. [CrossRef]

10 Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow. medRxiv [Preprint]. 2023 Feb 26:2023.02.21.23285886. Update in: J Med Internet Res. 2023 Aug 22;25:e48659. [CrossRef]

11 Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: An analysis of multiple clinical and research scenarios. J Med Syst. 2023;47(1):33. [CrossRef]