Original Article

ANATOLIAN
CURRENT MEDICAL

DOI: 10.38053/acmj.1649283

Anatolian Curr Med J. 2025;7(3):326-330

The role of Artificial Intelligence in Celiac disease support: analyzing ChatGPT's effectiveness for healthcare providers and patients

Yunus Halil Polat[1], Rasim Eren Cankurtaran[2]

[1]*Department of Gastroenterology, Ankara Training and Research Hospital, University of Health Sciences, Ankara, Turkiye*
[2]*Department of Gastroenterology, Ankara Etlik City Hospital, Ankara, Turkiye*

Cite this article as: Polat YH, Cankurtaran RE. The role of Artificial Intelligence in Celiac disease support: analyzing ChatGPT's effectiveness for healthcare providers and patients. *Anatolian Curr Med J.* 2025;7(3):326-330.

Received: 01.03.2025 ◆ **Accepted**: 28.04.2025 ◆ **Published**: 30.05.2025

ABSTRACT

Aims: Celiac disease is a significant autoimmune disorder that affects a substantial portion of the population, necessitating accurate and reliable information for effective management. Despite its importance, there are currently no studies assessing the performance of Artificial Intelligence (AI) language models, such as ChatGPT, in providing information on this condition. This study aims to evaluates the reliability and usefulness of ChatGPT's responses to frequently asked questions regarding celiac disease, thereby filling a critical gap in understanding the capabilities of AI in this important healthcare context.

Methods: A total of 20 questions (10 for patients/caregivers and 10 for healthcare professionals) were prepared based on the most frequently searched queries about celiac disease using Google Trends. Responses generated by ChatGPT and scored by two independent Likert raters.

Results: The analysis revealed strong inter-rater reliability, with Cronbach's alpha values of α=0.839 for reliability and α=0.753 for usefulness, indicating robust agreement between raters. Notably, the highest reliability and usefulness scores for the patient and caregiver group were associated with questions on symptoms, the celiac disease diet, and gluten-free products. For the healthcare professionals group, key topics included diagnosis, pathological classification, and celiac disease comorbidities. Importantly, no significant differences were found between raters in the evaluation of reliability (p=0.939) and usefulness (p=0.102).

Conclusion: This study demonstrates that AI-based language models like ChatGPT can serve as reliable and useful resources for both patients and healthcare professionals seeking information about celiac disease. While the model excelled in addressing commonly discussed topics, it revealed limitations in handling complex issues, emphasizing the need for ongoing refinement of AI tools. These findings support the integration of AI in healthcare communication, highlighting its potential to enhance access to crucial health information while underscoring the importance of continual improvement to meet diverse user needs.

Keywords: Celiac disease, chatGBT, artificial intelligence

INTRODUCTION

Celiac disease (CD) is a chronic autoimmune disorder triggered by the consumption of gluten-containing foods, leading to damage in the small intestine. There has been a significant increase in the incidence of CD over the last 50 years. This increase is due to the detection of more new cases as a result of improved diagnostic tools and widespread screening of people at high risk. However, the vast majority of patients with CD remain undetected worldwide.[1] In Western countries, the prevalence of CD is approximately 0.6% when confirmed histologically and about 1% based on serological screening of the general population.[2] The primary treatment for this condition is the lifelong adherence to a strict gluten-free diet. Gluten, a protein found in wheat, barley and rye, damages the villi in the small intestine of people with CD, affecting the absorption of nutrients.[3] Strict compliance with a gluten-free diet is essential for alleviating symptoms and preventing long-term complications associated with the disease.[4] However, managing a gluten-free diet can be challenging, particularly in terms of daily food choices. It is crucial that both celiac patients and healthcare professionals have access to accurate and reliable information regarding CD.

In recent years, advancements in artificial intelligence (AI) technologies, such as natural language processing (NLP) and machine learning, have revolutionized access to health information.[5] AI-based tools have the potential to support healthcare by providing rapid and personalized information to both patients and professionals.[6] Various chatbots, such as Woebot, Your.MD, HealthTap, Cancer Chatbot, VitaminBot,

Corresponding Author: Yunus Halil Polat, yunushpolat@gmail.com

This work is licensed under a Creative Commons Attribution 4.0 International License.

Babylon Health, Safedrugbot, and Ada Health, are being utilized for a range of functions in the healthcare industry.[7]

Chat Generative Pre-trained Transformer (ChatGPT), developed by OpenAI and released in November 2022, is one of the most advanced examples of large language models (LLMs). As a large language model, ChatGPT can provide human-like responses based on vast amounts of learned data from various sources.[8] However, the effectiveness and reliability of AI-based tools, especially in the context of providing medical information, remain areas of active research.[9]

In the field of gastroenterology, many studies investigating the reliability of ChatGPT for many diseases have been published.[10,11] According to our research, we did not find any study in the literature measuring the reliability and usefulness of ChatGPT in CD disease.

In this study, we aimed to evaluate the reliability and usefulness of ChatGPT in relation to CD, a significant chronic illness expected to gain even more importance. Through this research, we sought to shed light on whether ChatGPT is a reliable resource from the perspectives of both patients and healthcare professionals.

METHODS

The study adhered to the ethical standards outlined in the Helsinki Declaration and complied with national regulations in the respective field. Ethics committee approval was not required as the study did not involve the use of human or animal data.

For this study, we focused on CD and identified specific questions relevant to the condition. A total of 20 questions were selected, with the first 10 based on Google Trends searches to capture the most common questions from patients. The remaining 10 questions were developed according to the latest guidelines[1,12] for healthcare professionals. Separate Google Trends searches were conducted on 26 August 2024 to identify the top search terms related to CD. Search trends for relevant terms were analysed based on global data from 2004 to the present within the Health sub-category. In the results, the "most relevant" option was selected from the "relevant questions" section. This analysis revealed the most frequently searched keywords on Google for CD. Repetitive keywords with similar meanings were filtered out. Based on these keywords, questions were formulated covering various aspects of the disease, including its characteristics, causes, symptoms, treatment options and dietary considerations. The next ten questions were intended for healthcare professionals. These questions were developed by two gastroenterologists.

The questions were entered into the prompt section of the ChatGPT AI chatbot. As the conversation progressed, different users rephrased the next question in separate sessions. This method was employed to ensure that the response to each question was not influenced by previous questions or answers. Each response was recorded in a separate file. The answers provided by ChatGPT-4 were sourced from the premium version available on 14 March 2023.

Each chatbot was rated on a scale of 1-7 (1 being the lowest, 7 being the highest) in two categories for reliability and usefulness.[13] These scales are presented in our study with a slight modification in Table 1, 2. All responses were assessed by two independent gastroenterology experts who were blinded to each other's responses to avoid potential bias.

Table 1. Reliability score

1. Completely unsafe: None of the information provided can be verified from medical sources or contains inaccurate and incomplete information.
2. Very unsafe: Most of the information cannot be verified from medical sources or are partially correct but contains important incorrect or incomplete information.
3. Relatively reliable: The majority of the information provided are verified from medical scientific sources, but there are some important incorrect or incomplete information.
4. Reliable: Most of the information provided are verified from medical scientific sources, but there are some minor inaccurate or incomplete information.
5. Relatively very reliable: Most of the information provided are verified from medical scientific sources, and there is very little incorrect or incomplete information.
6. Very reliable: Most of the information provided are verified from medical scientific sources, and there is almost no inaccurate or incomplete information.
7. Absolutely reliable : All of the information provided are verified from medical scientific sources, and there is no inaccurate or incomplete information or missing information.

Table 2. Usefulness score

1. Not useful at all: Unintelligible language, contradictory information, and missing important information. Not useful for users.
2. Very little useful: Partly clear language is used. Some important information are missing or incorrect. For users, limited use is possible.
3. Relatively useful: Clear language is used. Most important information are mentioned, but some important information are incomplete or incorrect. Useful for users.
4. Partly useful: Clear language is used. Some important information are missing or incorrect, but most important information are addressed. Somewhat useful for users.
5. Moderately useful: Clear language is used and most important information are covered, but some important information are still incomplete or incorrect. Useful for users.
6. Very useful: Clear language is used. All important information are mentioned, but some unimportant information or details are also mentioned. Very useful for users.
7. Extremely useful: Clear language is used and all important information are mentioned. Extremely useful to users, and additional information and resources are also provided.

Statistical Analysis

The statistical analyses were performed using Statistical Package for the Social Sciences (SPSS 25.0 for Windows; IBM, Armonk, NY, USA) software package. The inter-rater compliance was assessed with Cronbach α and 95% confidence intervals (CI). According to intraclass correlation coefficient results, positive values ranging from 0 to 0.2 indicate poor agreement; 0.2 to 0.4 indicate fair agreement; 0.4 to 0.6 indicate moderate agreement; 0.6 to 0.8 indicate good agreement; and 0.8 to 1 indicate very good agreement. The variables were evaluated using the Shapiro-Wilk test to determine whether or not they exhibited a normal distribution. In descriptive statistics, the data were expressed as mean±standard deviation (SD). Independent t test was used to compare two groups difference and statistically significant difference among the groups was performed by analysis of variance test. The significance level for this study was set at p&lt;0.05.

RESULTS

In total, 20 different questions were presented to the OpenAI chatbot. The first ten questions were for CD patients and carers and the next ten questions were highly specific for medical professionals in terms of CD. The supplementary material of the study contains a comprehensive list of all questions and their corresponding answers.

We had two experts evaluate the responses given by ChatGPT on CD. The results of the evaluation for questions related to reliability and usefulness are shown in Table 3, 4 respectively.

Table 3. Distribution, comparison, and agreement of inter-rater reliability scores

	Rater#1	Rater#2	Cronbach's α (95% CI lower-upper)
1. What's CD?	5	5	0.789 (0.152-0.948)
2. Causes	5	5	
3. Symptoms	6	6	
4. CD diet	6	6	
5. Treatment	5	6	
6. Safe snacks	5	6	
7. Gluten free products	6	6	
8. Exercise and lifestyle	5	5	
9. Vitamin and suplements	6	6	
10. Alcohol consumption	6	6	
11. Diagnose	6	5	0.625 (0.510-0.907)
12. Pathological classification	5	5	
13. Association with other diseases	4	4	
14. Pregnancy	5	5	
15. Complications	5	4	
16. Genetic alleles	5	5	
17. Treatments	5	4	
18. CD and IBS	4	5	
19. CD and drugs	4	4	
20. CD and lactose intolerance	4	4	

CI: Confidence intervals, CD: Celiac disease, IBS: Irrıtable Bowel syndrome

Inter-rater Cronbach α values for reliability and usefulness total scores between raters showed good and very good agreement (α=0.839 and α=0.753, respectively).

The question topics were rated using Likert scores ranging from 3 to 7.

In terms of topics, the highest reliability score for patient and caregiver group was for both raters: point 6 (rater 1: symptoms, CD diet,gluten free products, vitamin and suplements, alcohol consumption ; rater 2: symptoms, CD diet, treatment, safe snacks, gluten free products, vitamin and suplements, alcohol consumption) and the highest reliability score for proffesionals group was for rater 1: point 6 and was for rater 2: point 5 (rater 1: diagnose ; rater: 2: diagnose, pathological classification, pregnancy, genetic alleles, CD and IBS). The highest usefullness score for patient and caregiver group was for both raters: point 7 (rater 1: symptoms, CD diet,gluten free products, vitamin and suplements; rater 2: symptoms) and the highest usefullness score for proffesionals group was for both raters point 6 (rater 1: diagnose, pathological classification, pregnancy, genetic alleles, CD and IBS; rater: 2: diagnose, pregnancy, complications, treatments, CD and IBS).

Table 4. Distribution, comparison, and agreement of inter-rater usefulness scores

	Rater#1	Rater#2	Cronbach's α (95% CI lower-upper)
1. What's CD?	6	6	0.691 (0.243-0.923)
2. Causes	6	5	
3. Symptoms	7	7	
4. CD diet	7	6	
5. Treatment	6	5	
6. Safe snacks	6	6	
7. Gluten free products	7	6	
8. Exercise and lifestyle	6	6	
9. Vitamin and suplements	7	6	
10. Alcohol consumption	6	6	
11. Diagnose	6	6	0.640 (0.449-0.911)
12. Pathological classification	6	5	
13. Association with other diseases	5	5	
14. Pregnancy	6	6	
15. Complications	5	6	
16. Genetic alleles	6	5	
17. Treatments	6	6	
18. CD and IBS	6	6	
19. CD and drugs	5	5	
20. CD and lactose intolerance	5	4	

CI: Confidence intervals, CD: Celiac disease, IBS: Irrıtable Bowel syndrome

In terms of topics, the lowest reliability score for patient and caregiver group was for both raters point 5 (rater 1: what's CD?, causes, treatment, safe snacks, exercise and lifestyle ; rater 2: what's CD?, causes, exercise and lifestyle) and the lowest reliability score for proffesionals group was for both raters point 4 (rater 1: association with other diseases, CD and IBS, CD and drugs, CD and lactose intolerance ; rater: 2: association with other diseases, complications, treatments, CD and drugs, CD and lactose intolerance). The lowest usefullness score for patient and caregiver group was for rater 1 point 6 was for rater 2 point 5 (rater 1: what's CD?, causes, treatment, safe snacks, exercise and lifestyle, alcohol consumption ; rater 2: causes, treatment) and the lowest usefullness score for proffesionals group was for rater 1 point 5 was for rater 2 point 4 (rater 1: association with other diseases, complications, CD and drugs, CD and lactose intolerance ; rater: 2: CD and lactose intolerance).

The total scores of the topics and their evaluation by each rater are shown in Table 5 and Figure. There was no significant difference between the reliability (rater#1 mean:5.10±0.71, rater#2 mean:5.10±0.78) and usefulness (rater#1 mean:6.0±0.64, rater#2 mean:5.65±0.67) total scores of both raters (p=0.939 and p=0.102, respectively).

Table 5. Comparison of the reliability and usefulnes total scores of resources according to patients and professonals

	For patients	For professionals	p
Rater #1			
Reliability	5.50±0.52	4.70±0.67	0.008
Usefulness	6.40±0.51	5.60±0.51	0.003
Rater #2			
Reliability	5.70±0.48	4.50±0.52	0.001
Usefulness	5.90±0.56	5.40±0.69	0.049
Reliability p	0.388	0.470	
Usefulness p	0.054	0.476	

*Mean±standard deviation, independent t test

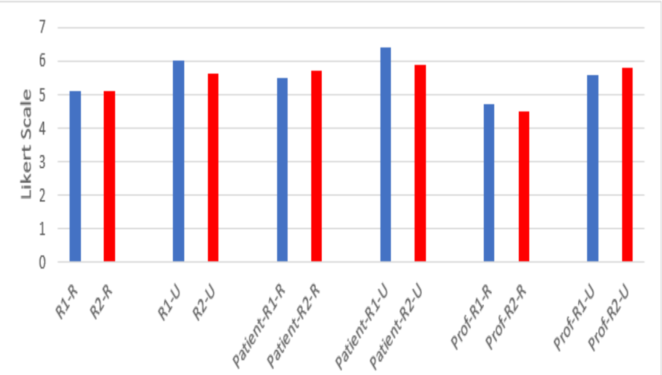

Figure. Total reliability and usefullness scores and scores according to patients, professionals and raters distribution

R: Reliability, U: Usefullness, Prof: Professionals

DISCUSSION

This study evaluated the reliability and usefulness of responses provided by an AI-based language model, ChatGPT, to common questions about CD posed by two distinct groups: patients and caregivers, and healthcare professionals. The results demonstrate that ChatGPT's responses were rated with high inter-rater reliability and usefulness, as evidenced by Cronbach's alpha values of 0.839 for reliability and 0.753 for usefulness. These findings indicate a good to very good level of agreement between raters, confirming that the AI model consistently provided information that was both accurate and relevant.

Reliability and Usefulness of Responses

The analysis showed that the highest reliability and usefulness scores for the patient and caregiver group were consistently awarded for responses addressing common topics such as symptoms, the CD diet, gluten-free products, and the role of vitamins and supplements. This is an indication that ChatGPT was effective in addressing topics that are frequently discussed in public health resources and are central to patient education.[4,14] In contrast, the lowest reliability scores were observed for more general questions, such as "What is CD?" and questions about causes and lifestyle, reflecting potential gaps in how the AI model synthesizes more foundational or lifestyle-related information. This is consistent with previous studies highlighting AI models' variable performance when addressing nuanced or highly specialized information.[15]

For the healthcare professionals group, the highest reliability scores were seen in responses related to diagnostic processes, pathological classification, and the interplay between CD and other gastrointestinal conditions, such as irritable bowel syndrome (IBS). This underscores ChatGPT's strength in addressing medically technical content, likely because such topics are well-documented and discussed in the medical literature.[16] However, reliability and usefulness dropped when the topics involved complex, multi-faceted issues, such as the association between CD and other diseases or lactose intolerance. This trend might reflect the inherent limitations of current AI models in dealing with interdisciplinary or comorbid conditions, which often require a more integrative understanding of disease pathophysiology.[17]

Inter-Rater Consistency

The absence of a significant difference between raters regarding both reliability and usefulness (p=0.939 and p=0.102, respectively) suggests that the AI model's performance was consistent across evaluators with different expertise. This finding is important, as it indicates that the content generated by ChatGPT is perceived similarly by healthcare professionals, regardless of individual biases or interpretative differences. High Cronbach alpha values further reinforce the robustness of these assessments, aligning with previous studies on the reliability of AI-generated content in healthcare settings.[18,19]

Limitations and Future Directions

Despite the promising results, several limitations should be acknowledged. Firstly, while the AI model performed well on well-defined and frequently discussed topics, it showed limitations in providing comprehensive answers to more complex questions, particularly those involving comorbidities or nuanced lifestyle modifications. Future versions of AI tools may benefit from incorporating more diverse training datasets that include interdisciplinary research and guidelines for managing chronic diseases like CD with associated conditions.

Moreover, this study focused on the evaluation of AI responses in the context of CD, a specific and relatively well-documented condition. Generalizability to other chronic diseases remains to be studied. Future research should also examine the impact of AI-generated information on patient outcomes, including how well patients adhere to medical advice received from such tools.

In addition to these limitations, it is important to highlight the ethical considerations associated with the use of AI in healthcare settings. Issues such as accountability for incorrect or harmful advice, the protection of patient privacy, and the necessity for informed consent regarding the use of AI-generated content remain areas of ongoing debate. The absence of clear legal frameworks regulating the medical application of AI systems raises questions about liability, particularly in scenarios where AI-generated suggestions may contribute to adverse patient outcomes. Given these uncertainties, AI-based tools like ChatGPT should be viewed strictly as supportive educational resources rather than replacements for professional medical judgment.

Furthermore, while AI models can efficiently synthesize well-documented medical knowledge, their use should always be supervised by qualified healthcare professionals, especially when clinical decision-making is involved. Without appropriate oversight, there is a risk that users—whether patients or providers—might misinterpret AI outputs as authoritative medical advice, which could inadvertently affect treatment decisions or health behaviors. These considerations emphasize the need for cautious integration of AI technologies into healthcare, ensuring that their implementation remains ethically sound and clinically responsible.

CONCLUSION

This study provides valuable insights into the potential of AI-based tools, such as ChatGPT, to serve as reliable and useful resources for both patients and healthcare professionals in the context of CD. While AI models are still evolving, they hold significant promise in enhancing access to health information, especially when used to complement traditional healthcare resources. Continued research is necessary to refine these systems and ensure they meet the evolving needs of patients and healthcare providers alike.

ETHICAL DECLARATIONS

Ethics Committee Approval

Since the study did not involve the use of human or animal data, ethics committee approval was not necessary.

Informed Consent

Since the study did not involve the use of human data, informed consent was not necessary.

Referee Evaluation Process

Externally peer-reviewed.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Financial Disclosure

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

Author Contributions

All of the authors declare that they have all participated in the design, execution, and analysis of the paper, and that they have approved the final version.

REFERENCES

1. Al-Toma A, Volta U, Auricchio R, et al. European Society for the Study of Coeliac Disease (ESsCD) guideline for coeliac disease and other gluten-related disorders. *United European Gastroenterol J.* 2019;7(5):583-613. doi:10.1177/2050640619844125
2. Fasano A, Berti I, Gerarduzzi T, et al. Prevalence of celiac disease in at-risk and not-at-risk groups in the United States: a large multicenter study. *Arch Intern Med.* 2003;163(3):286-292. doi: 10.1001/archinte.163.3.286
3. Green PH, Cellier C. Celiac disease. *N Engl J Med.* 2007;357(17):1731-1743. doi:10.1056/NEJMra071600
4. Rubio-Tapia A, Hill ID, Kelly CP, Calderwood AH, Murray JA; American College of Gastroenterology. ACG clinical guidelines: diagnosis and management of celiac disease. *Am J Gastroenterol.* 2013;108(5):656-576; quiz 677. doi:10.1038/ajg.2013.79
5. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019;25(1):44-56. doi:10.1038/s41591-018-0300-7
6. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* 2017;542(7639):115-118. doi:10.1038/nature21056
7. Palanica A, Flaschner P, Thommandram A, Li M, Fossat Y. Physicians' perceptions of chatbots in health care: cross-sectional web-based survey. *J Med Internet Res.* 2019;21(4):e12887. doi:10.2196/12887
8. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. arXiv preprint arXiv:2005.14165 (2020).
9. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol.* 2017;2(4):230-243. doi:10.1136/svn-2017-000101
10. Kerbage A, Kassab J, El Dahdah J, et al. Accuracy of ChatGPT in common gastrointestinal diseases: impact for patients and providers. *Clin Gastroenterol Hepatol.* 2024;22(6):1323-1325.e3. doi:10.1016/j.cgh.2023.11.008
11. Klang E, Sourosh A, Nadkarni GN, Sharif K, Lahat A. Evaluating the role of ChatGPT in gastroenterology: a comprehensive systematic review of applications, benefits, and limitations. *Therap Adv Gastroenterol.* 2023;16:17562848231218618. doi:10.1177/17562848231218618
12. Rubio-Tapia A, Hill ID, Semrad C, et al. American College of gastroenterology guidelines update: diagnosis and management of celiac disease. *Am J Gastroenterol.* 2023;118(1):59-76. doi:10.14309/ajg.0000000000002075
13. Uz C, Umay E. "Dr ChatGPT": is it a reliable and useful source for common rheumatic diseases? *Int J Rheum Dis.* 2023;26(7):1343-1349. doi:10.1111/1756-185X.14749
14. Green PH, Cellier C. Celiac disease. *N Engl J Med.* 2007;357(17):1731-1743. doi:10.1056/NEJMra071600
15. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019;25(1):44-56. doi:10.1038/s41591-018-0300-7
16. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* 2017;542(7639):115-118. doi:10.1038/nature21056
17. Bot B, Suver C, Niederhoffer K. The use of artificial intelligence in medical literature review: a novel approach to evidence-based medicine. *Med Hypotheses.* 2017;105:1-3. doi:10.1016/j.mehy.2017.05.015
18. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol.* 2017;2(4):230-243. doi:10.1136/svn-2017-000101
19. Nuti SV, Wayda B, Ranasinghe I, et al. The use of google trends in health care research: a systematic review. *PLoS One.* 2014;9(10):e109583. doi:10.1371/journal.pone.0109583