Journal of Crohn's and Colitis, 2024, **XX**, 1–7
https://doi.org/10.1093/ecco-jcc/jjae040
Advance access publication 23 March 2024
Original Article

Accuracy of Information given by ChatGPT for Patients with Inflammatory Bowel Disease in Relation to ECCO Guidelines

Martina Sciberras,[a] Yvette Farrugia,[a] Hannah Gordon,[b,c] Federica Furfaro,[d] Mariangela Allocca,[d] Joana Torres,[e,f,g] Naila Arebi,[h,i], Gionata Fiorino,[d,j] Marietta Iacucci,[k] Bram Verstockt,[l,m] Fernando Magro,[n] Kostas Katsanos,[o] Josef Busuttil,[p] Katya De Giovanni,[p] Valerie Anne Fenech,[a] Stefania Chetcuti Zammit,[a] Pierre Ellul[a]

[a]Department of Medicine, Division of Gastroenterology, Mater Dei Hospital, Msida, Malta
[b]Department of Gastroenterology, Barts Health NHS Trust, London, UK
[c]Translational Gastroenterology and Liver Unit, John Radcliffe Hospital, University of Oxford, Oxford, UK
[d]IRCCS OSPEDALE San Raffaele, Gastroenterology and Endoscopy, IBD Center, Milan, Italy
[e]Division of Gastroenterology, Hospital da Luz, Lisbon, Portugal
[f]Division of Gastroenterology, Hospital Beatriz Ângelo, Loures, Portugal
[g]Faculdade de Medicina, Universidade de Lisboa, Lisbon, Portugal
[h]Department of Inflammatory Bowel Disease, St Mark's National Bowel Hospital, London, UK
[i]Department of Metabolism, Digestion and Reproduction, Imperial College London, London, UK
[i]IBD Unit, San Camillo-Forlanini Hospital, Rome, Italy
[k]APC Microbiome Ireland, College of Medicine and Health, University College of Cork, Cork, Ireland
[l]Department of Gastroenterology and Hepatology, University Hospitals Leuven, KU Leuven, Leuven, Belgium
[m]Department of Chronic Diseases and Metabolism, KU Leuven, Leuven, Belgium
[n]CINTESIS@RISE, Faculty of Medicine of the University of Porto, Porto, Portugal
[o]Division of Gastroenterology, Department of Internal Medicine, Faculty of Medicine, University of Ioannina School of Health Sciences, Ioannina, Greece
[p]Association for Crohn`s and Colitis, Malta

Corresponding author: Martina Sciberras, Department of Medicine, Division of Gastroenterology, Mater Dei Hospital, Msida, Malta. Email: martina.sciberras.2@gov.mt

Abstract

Background: As acceptance of artificial intelligence [AI] platforms increases, more patients will consider these tools as sources of information. The ChatGPT architecture utilizes a neural network to process natural language, thus generating responses based on the context of input text. The accuracy and completeness of ChatGPT3.5 in the context of inflammatory bowel disease [IBD] remains unclear.

Methods: In this prospective study, 38 questions worded by IBD patients were inputted into ChatGPT3.5. The following topics were covered: [1] Crohn's disease [CD], ulcerative colitis [UC], and malignancy; [2] maternal medicine; [3] infection and vaccination; and [4] complementary medicine. Responses given by ChatGPT were assessed for accuracy [1—completely incorrect to 5—completely correct] and completeness [3-point Likert scale; range 1—incomplete to 3—complete] by 14 expert gastroenterologists, in comparison with relevant ECCO guidelines.

Results: In terms of accuracy, most replies [84.2%] had a median score of ≥4 (interquartile range [IQR]: 2) and a mean score of 3.87 [SD: ±0.6]. For completeness, 34.2% of the replies had a median score of 3 and 55.3% had a median score of between 2 and <3. Overall, the mean rating was 2.24 [SD: ±0.4, median: 2, IQR: 1]. Though groups 3 and 4 had a higher mean for both accuracy and completeness, there was no significant scoring variation between the four question groups [Kruskal–Wallis test $p > 0.05$]. However, statistical analysis for the different individual questions revealed a significant difference for both accuracy [$p < 0.001$] and completeness [$p < 0.001$]. The questions which rated the highest for both accuracy and completeness were related to smoking, while the lowest rating was related to screening for malignancy and vaccinations especially in the context of immunosuppression and family planning.

Conclusion: This is the first study to demonstrate the capability of an AI-based system to provide accurate and comprehensive answers to real-world patient queries in IBD. AI systems may serve as a useful adjunct for patients, in addition to standard of care in clinics and validated patient information resources. However, responses in specialist areas may deviate from evidence-based guidance and the replies need to give more firm advice.

Key Words: Artificial intelligence; inflammatory bowel disease; health communication; patient education

1. Introduction

Traditionally, patients depended upon physicians for medical information. However, retainment of medical information given in an outpatient clinic setting is limited. To counteract this, patient information leaflets and patient support groups were introduced.[1]

The availability of the Internet as a source of health information has further changed the way individuals seek medical knowledge. Recognized associations such as the European Federation for Crohn's and Colitis Associations [EFCCA] provide readily available and trustworthy information online.[2] Patients can also use any available search engine and obtain medical information. However, the reliability of such online resources remains a concern, with misleading information widespread.[3,4]

A survey conducted in 2020 revealed that 55% of European Union [EU] citizens aged 16–74 years use online health and disease information, demonstrating the demand for digital resources.[5] In recent years, artificial intelligence [AI] has made significant advancements. The public gained its first easy access to AI when OpenAI released Chat Generative Pre-Trained Transformer [ChatGPT] [OpenAI, L.L.C., San Francisco, CA, USA], a large language model, in November 2022.[6] Utilizing natural language processing, ChatGPT has the ability to understand text and simulate human-like responses. This novel platform holds immense potential and merits a thorough study of its capabilities and limitations in the context of public health.[7] The ChatGPT program is able to understand and generate responses using a text-based interface and is based on the generative pre-trained transformer [GPT] architecture. The GPT architecture utilizes a neural network to process natural language, thus generating responses based on the context of input text.[8]

As acceptance of AI platforms increases, more patients will consider these tools as sources of information. Nevertheless, the trustworthiness of online resources remains a barrier. The potential of ChatGPT in public health has been recognized in the literature.[7] It is essential to consider both the advantages as well as the limitations of this platform, particularly regarding its accuracy. Studies have started to investigate the performance of AI systems and their ability to provide accurate answers to medical questions. In 2023, Johnson *et al*. demonstrated its potential in providing accurate medical information 3 months after its launch. Most of the answers generated by the AI chatbot [57.8%] were rated as 'nearly correct' or 'correct' by the physicians. Furthermore, ChatGPT provided comprehensive answers in a significant proportion of cases [53.5%].[9] Lee *et al*. compared the answers provided by ChatGPT regarding colonoscopies to the answers available on patient information sites, such as hospital webpages. Both AI-generated and non-AI-generated answers exhibited similar content and accuracy. However, the answers provided by ChatGPT were found to be easier to understand.[10] ChatGPT has also demonstrated competence by performing at the passing threshold for the United States Medical Licensing Exam [USMLE] Step 1 and Step 2. It answered correctly over 60% of questions, which is commonly considered the passing standard.[11]

However, in another study both ChatGPT-3 and GPT-4 did not pass the American College of Gastroenterology self-assessment test.[12]

The aim of the present study is to assess the reliability of responses generated by ChatGPT for a set of frequently asked questions posed by patients with inflammatory bowel disease [IBD].

2. Methodology

This is a prospective study whereby a series of commonly asked questions by patients in the clinic were entered into ChatGPT. These questions were given to us by two patient representatives from the National Association of Crohn's and Colitis. Patients can submit medical inquiries to the Maltese Association for Crohn's and Colitis, which are subsequently forwarded to medical specialists. These questions, coupled with feedback from other members of the Association, served as valuable inputs for the questions used in this study. After receiving the questions, the coordinators organized them into distinct subgroups for ease of reference. The 38 questions and answers were subdivided into four categories, based on European Crohn's and Colitis Organization [ECCO] guidelines and topical reviews, these being:

Group 1: Ulcerative colitis, Crohn's disease and Malignancy Guidelines. Questions 1–10
Group 2: Pregnancy and Fertility. Questions. 11–21
Group 3: Vaccinations and Infections Questions. 22–28
Group 4: Herbal Options and Complementary Medicine. Questions. 29–38

These questions were then inputted into ChatGPT and the replies were assessed. To ensure consistency, all questions were entered into the ChatGPT engine by one investigator, using the freely available GPT-3.5-based version. All the doctors doing the analysis worked in European countries. Questions and answers can be found in full in the Supplementary Material.

These replies were then analysed by 14 gastroenterologists with a special interest in IBD, with most of them being involved in the ECCO guidelines and all of them being involved in National/International guidelines. The answers were assessed on accuracy using a 5-point Likert scale and graded as follows:

1. Completely incorrect
2. More incorrect than correct [>75% incorrect]
3. Approximately equal correct and incorrect
4. More correct than incorrect [>75% correct]
5. Completely correct

The answers were also assessed on completeness using a 3-point Likert Scale as follows:

1. Incomplete
2. 50% complete
3. Complete

Score results were listed descriptively [median, mean, interquartile range, standard deviation], and were compared between groups using non-parametric testing. Statistical analysis was performed using SPSS v27. The Friedman test was used to compare mean rating scores between several related statements. One-way ANOVA was used to compare means between the different groups.

3. Results

3.1. General comments on answers provided by ChatGPT

3.1.1. Group 1: general characteristics of CD, UC and malignancy

The comments by the reviewers for this group [which had the second lowest mean for accuracy and lowest mean for

completeness; Table 1; Figures 1 and 2] related to facts that very often in the clinic one would usually be either aware of or would ask the patient. For example, one would be aware of the presence or absence of more medical information such as the presence or absence of primary sclerosing cholangitis and/or post-inflammatory polyps during colonoscopy. The latter would alter the colonoscopy screening interval. This medical information would usually help the clinician in providing better answers. Furthermore, its inability to give exact surgical or malignancy risks or recommend screening tests and/or tools was another significant limitation in the answers that were provided. Within this group, ChatGPT was unable to give a comprehensive reply on available screening tests for patients on immunosuppressants and very often the data on endoscopic surveillance were imprecise.

3.1.2. Group 2: pregnancy and fertility

The main comments pertaining to Group 2 related mostly to the information given on medications, active disease versus remission, and their relation to fertility. Active disease is usually the greatest hinderance in achieving a successful pregnancy, and this fact did not come out well in the replies. This group had the lowest overall mean score for completeness and second lowest for accuracy [Table 1; Figures 1 and 2]. Furthermore two questions in this group were among the three which scored lowest for both completeness and accuracy. In one question, the main concern was that the important distinction between live and non-live vaccines was not done, especially since this was in the important context of immunosuppression. The other answer with a low score did not address the question and did not pick up the fact that the patient was male and not female.

3.1.3. Group 3: vaccinations and infections

This group had higher means for both accuracy [4.0] and completeness [2.3]. The main comments throughout were that at times the answers provided could not correlate well live and non-live vaccines with the various immunosuppressive medications.

3.1.4. Group 4: herbal options and complementary medicine

Group 4 had the highest means for both accuracy [4.23] and completeness [2.44]. The questions which scored lowest related to nutrition, as the answers were vague, especially those that relate to ulcerative colitis.

Some other comments were common for most ChatGPT answers across all the different specialists and topics, these being that at times the replies were vague and did not assess

Table 1. Question comparison per group

	Group 1	Group 2	Group 3	Group 4
Mean accuracy [5-point LS]	3.67	3.65	4.00	4.23
Median accuracy [5-point LS]	4	4	4	4
Mean completeness [3-point LS]	2.05	2.18	2.30	2.44
Median completeness [3-point LS]	2	2	3	2.75

LS: Likert scale.

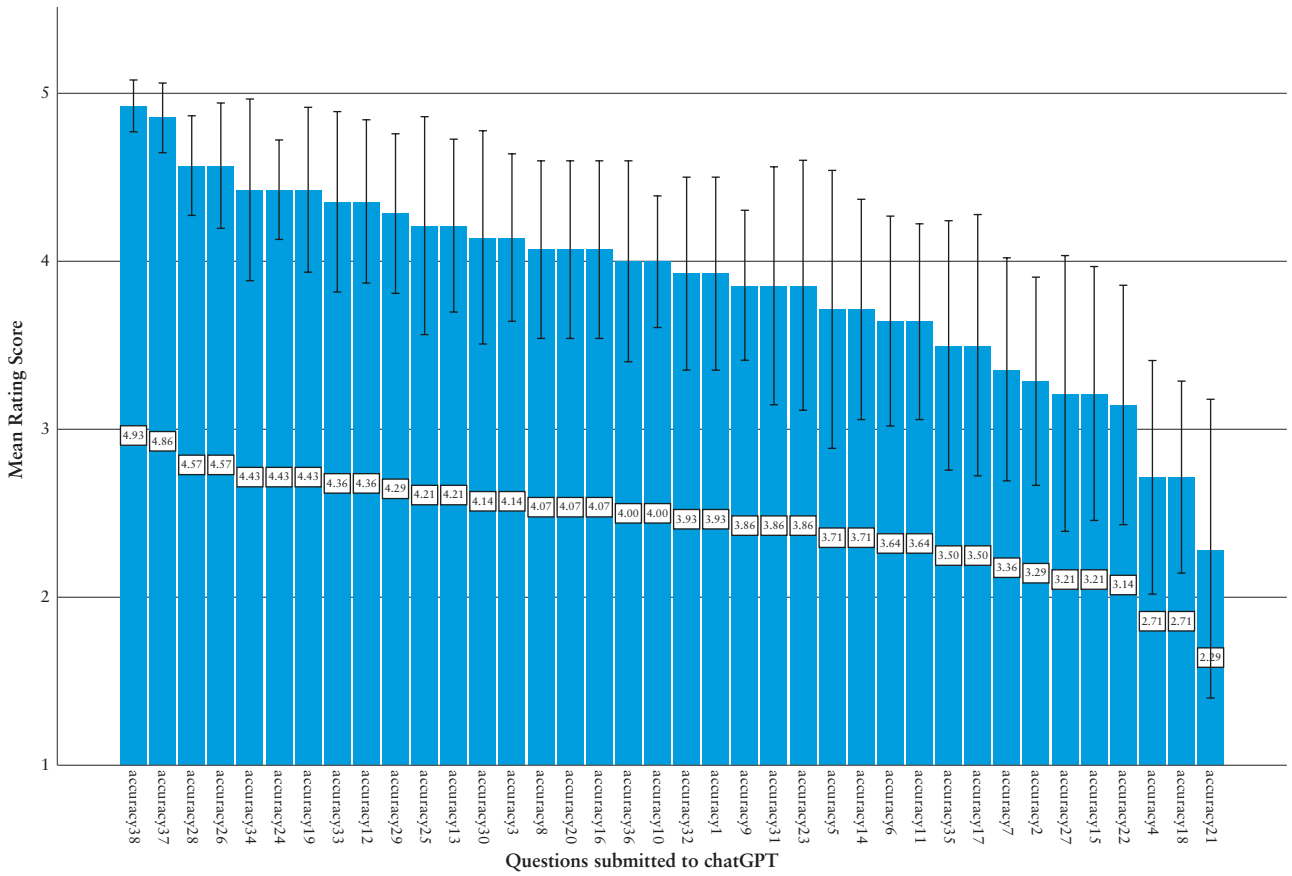

Figure 1. Mean score rating per question for accuracy.

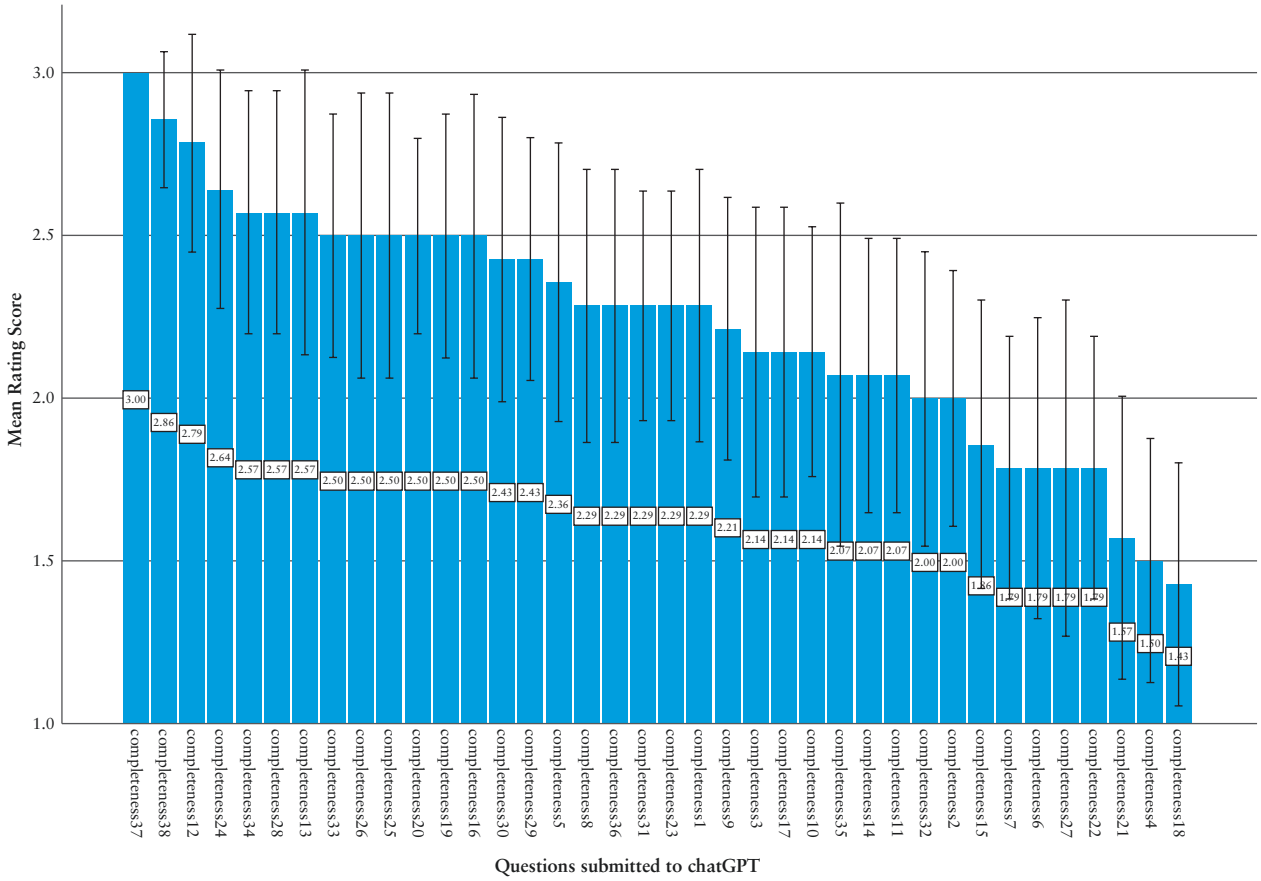

Figure 2. Mean score rating per question for completeness.

specifically the answer to the question, the information that was given was broad information and not specific, and ultimately there was always the disclaimer that refers the patient to the physician for assessment.

3.2. Data analysis

3.2.1. Accuracy

The mean accuracy rating on the 5-point Likert scale was 3.87 (SD: ±0.6, median: 4, interquartile range [IQR]: 2) with 50.0% being ≥4 and 42.1% having a mean score between 3 and <4. Only three questions [7.9%] had a mean score <3, but still above 1 [mean 2.57]. In terms of accuracy, most replies [84.2%] had a median score of ≥4. Only 7.9% had a median score of <3, with the rest [92.1%] having a median score of between 3 and <4 [Table 2].

3.2.2. Completeness

For the 3-point Likert scale for completeness, the mean rating was 2.24 [SD: ±0.4, median: 2, IQR: 1] with 78.9% of means for each question being ≥2. In terms of completeness, 34.2% of the replies had a median score of 3, 55.3% had a median score of between 2 and <3, and the rest [10.5%] had a median score <2 [Table 2].

The three questions which had both a mean and median score <3 for accuracy had a mean and median score of <2 for completeness.

3.3. Statistical analysis

The results were analysed for each section and for each reviewer. This analysis did not demonstrate any significant scoring variation for each of the four sections for each of the reviewers [Kruskal–Wallis test $p > 0.05$].

When analysing the means for accuracy, there was a statistically significant difference between the reply ratings, with some items having significantly higher values than others [Friedman test $p < 0.001$; Figure 1]

The questions which rated the highest for accuracy were on smoking and additional medications while the questions which rated the lowest were on genetic risk to a patient's children and vaccinations if the parent patient is taking biologics.

When analysing the means for completeness, there was a statistically significant difference between the reply ratings, with some answers having significantly higher values than others [$p < 0.001$; Figure 2].

As seen in both Figures 1 and 2, answers 4, 18, and 21 rated the lowest in both accuracy and completeness. Using an independent samples t-test and comparing the means of the questions with the lowest mean scores [Questions 4, 18, and 21] with the other questions, there was a statistically significant difference for completeness [$t = 4.38$, $p < 0.001$, SD ±0.18] and accuracy [$t = 5.22$, $p < 0.001$, SD ±0.27].

In terms of the overall 38 questions, those which rated the highest for both accuracy and completeness were related to smoking while the lowest ones related to screening for malignancy, vaccinations, and family planning.

When comparing the means of the four groups, there was trend towards a difference in accuracy and completeness which approached statistical significance [$p = 0.07$ and 0.10 respectively]. When undertaking direct comparisons between groups, there was lower mean accuracy in Group

Table 2. Mean and median for accuracy and completeness for each question

Question	Mean accuracy [SD]	Median accuracy [IQR]	Mean completeness [SD]	Median completeness [IQR]
1	3.92 [±0.50	4 [1]	2.29 [±0.36]	2 [2]
2	3.29 [±0.54]	4 [2]	2 [±0.34]	2 [0]
3	4.14 [±0.44]	4 [1]	2.14 [±0.39]	2 [1]
4	2.71 [±0.61]	2.5 [2]	1.5 [±0.33]	1 [1]
5	3.71 [±0.73]	4 [2]	2.36 [±0.38]	2.5 [1]
6	3.64 [±0.55]	4 [1]	1.79 [±0.41]	2 [2]
7	3.36 [±0.58]	3 [2]	1.79 [±0.35]	2 [1]
8	4.07 [±0.46]	4 [1]	2.29 [±0.36]	2 [1]
9	3.86 [±0.39]	4 [0]	2.21 [±0.35]	2 [1]
10	4 [±0.34]	4 [0]	2.14 [±0.34]	2 [1]
11	3.64 [±0.51]	4 [1]	2.07 [±0.36]	2 [2]
12	4.36 [±0.43]	4.5 [1]	2.79 [±0.29]	3 [0]
13	4.21 [±0.45]	46 [1]	2.57 [±0.38	3 [1]
14	3.71 [±0.57]	4 [2]	2.07 [±0.37]	2 [1]
15	3.21 [±0.66]	4 [2]	1.86 [±0.39]	2 [1]
16	4.07 [±0.46]	4 [1]	2.5 [±0.38]	3 [1]
17	3.5 [±0.68]	4 [1]	2.14 [±0.39]	2[1]
18	2.71 [±0.50]	2 [1]	1.43 [±0.33]	1 [1]
19	4.43 [±0.43]	5 [1]	2.5 [±0.33]	3 [1]
20	4.07 [±0.46]	4 [1]	2.5 [±0.26]	2.5 [1]
21	2.29 [±0.78]	2 [3]	1.57 [±0.38]	1 [1]
22	3.14 [±0.62]	3 [1]	1.79 [±0.35]	2 [1]
23	3.86 [±0.65]	4 [1]	2.29 [±0.31]	2 [1]
24	4.43 [±0.26]	4 [1]	2.64 [±0.32]	3 [1]
25	4.21 [±0.56]	4.5 [1]	2.5 [±0.38]	3 [1]
26	4.57 [±0.33]	5 [1]	2.5 [±0.38]	3 [1]
27	3.21 [±0.71]	3.5 [2]	1.79 [±0.35]	1.5 [2]
28	4.57 [±0.26]	5 [1]	2.57 [±0.33]	3 [1]
29	4.29 [±0.42]	4 [1]	2.43 [±0.33]	2.5 [1]
30	4.14 [±0.56]	4 [1]	2.43 [±0.33]	3 [1]
31	3.86 [±0.62]	4 [2]	2.29 [±0.31]	2 [1]
32	3.93 [±0.50]	4 [2]	2 [±0.40]	2 [2]
33	4.36 [±0.47]	5 [1]	2.5 [±0.33]	3 [1]
34	4.43 [±0.47]	5 [1]	2.57 [±0.33]	3 [1]
35	3.5 [±0.65]	4 [3]	2.07 [±0.46]	2 [2]
36	4 [±0.52]	4 [1]	2.29 [±0.37]	2 [1]
37	4.86 [±0.18]	5 [0]	3 [±0.00]	3 [0]
38	4.93 [±0.14]	5 [0]	2.86 [±0.18]	3 [0]

SD: standard deviation; IQR: interquartile range.

1 [mean: 3.67] than Group 4 [mean: 4.23] [$p = 0.05$] and a significantly lower mean accuracy in completeness between Group 1 [mean: 2.04] and Group 4 [mean: 2.44] [$p = 0.038$; Table 1].

4. Discussion

This study in IBD was able to demonstrate that AI platforms, in this case ChatGPT-3.5, have promise for providing accurate and comprehensive medical information.

In our study, we analysed both accuracy and completeness of the replies in relation to published guidelines. In terms of accuracy, the majority of replies [84.2%] had a median score of ≥4 [IQR:2] with a mean accuracy rating of 3.87 [SD: ±0.6]. In terms of completeness, the mean rating was 2.24 [SD: ±0.4, median: 2, IQR: 1] with 78.9% of means for each question being ≥2. Overall, accuracy was high across all question groups. The questions which rated the highest for accuracy were on smoking and additional medications while the questions which rated the lowest related to risk of IBD inheritance and vaccinations. Generally, in the clinic, these might also be the more difficult questions for the clinician to answer. In our study we also analysed assessor variability across the questions groups, and there was none.

When analysing the data, the median accuracy scores were generally higher than mean scores, which reflected multiple instances where the chatbot was wrong. There was no statistical difference overall between the four groups of questions.

In 2023, Johnson *et al*. re-scored answers provided by ChatGPT ~1–2 weeks after the initial responses were generated. The median accuracy score of the original low-quality answers improved when they were re-generated. This improvement could be attributed to the continuous update and refinement of the model's algorithms and parameters and the impact of repetitive user feedback through reinforced learning.[9] We did not re-generate the answers given the ever-changing medical knowledge and updates which should reflect what happens in real life. However, this demonstrates the ability of AI models to improve in performance rapidly when reinforced using model refinement and thus the results provided could be better within a relatively short period of time.

One of the strengths of ChatGPT is its ability to go through massive amounts of information and produce responses in a manner that is conversational and easy to understand. As demonstrated by Johnson *et al*.[11] the content is also updated far more frequently than hospital-based patient information leaflets and other conventional sources of information. Furthermore, results from current search engines can be overwhelming for both patients and physicians and can be further complicated with the presence of irrelevant or misleading information.

While ChatGPT has shown promise in producing mostly accurate and comprehensive responses, instances of either incomplete answers with no recommendations or incorrect answers have been noted. Most of the comments were for those answers whose mean score for completeness was <3. These mostly related to incomplete screening guidance [skin, cervical, and endoscopic] and inability to distinguish between live and non-live vaccines especially in the context of immunosuppression.

In the fertility group, a question was asked by a male patient who was receiving azathioprine and infliximab. The AI did not pick up the fact that the patient was male, and the reply given was that as if he had been a female patient. Furthermore, it also gave advice that azathioprine is teratogenic during pregnancy and for infliximab it stated 'there is limited data from human studies'. In two other instances, the AI ignored questions relating to the small bowel and replied if there had been colonic rather than small bowel involvement. Another concern, which should be considered as basic medicine, was its inability to suggest a DEXA scan in patients on long-term corticosteroid therapy. One potential strategy to enhance the reliability and personalization of information on these platforms could involve implementing a minimum requirement for medical information input from users. This could be facilitated through checkboxes or dropdown menus, ensuring a more robust and tailored exchange of information.

In the study by Johnson *et al*., 8.3% of questions were assessed as completely incorrect, most often observed in answers rated as difficult by the physicians.[9] A similar study by Cao *et al*. found that when ChatGPT was asked a set of questions regarding hepatocellular carcinoma [HCC] surveillance and diagnosis, 25% of the answers were considered inaccurate and ChatGPT provided contradictory answers in some cases.[12] An additional limitation demonstrated by a 2023 review noted an inability to appreciate regional variations in medical guidelines in the context of the interval and indications for HCC screening.[13]

In our study, common comments by the different specialists were that at times the replies were vague and not assertive enough, not specific, and without any firm recommendation. Ultimately there was always the disclaimer to discuss with your healthcare professional. Though this ensures optimal medical care, it also further raises the question of why one should log into such sites when ultimately you are referred back to your clinician. One's expectation would be to have an accurate specific answer to their query with references for the evidence. None of the answers contained links to the source of evidence to support the recommendations.

Though ChatGPT-3.5 is available free of charge, a more advanced version, GPT-4, is available to paid subscribers. Having reliable medical information as free may enhance the global health of the IBD patients as financial difficulty has been associated with poorer health outcomes.[8,14] It may also facilitate and empower patients in having optimal medical care.

ChatGPT has some limitations. Its current training extends only to information up until 2021. However, it is likely that such resources can be very easily updated, and it is also highly unlikely that patient information or websites are going to be updated more regularly. Second, the quality and accuracy of the dataset utilized to train such platforms is unknown and there are no references to ensure quality control. It has also been demonstrated that ChatGPT is able to generate incorrect content that appears plausible from a scientific point of view.[15]

Recently a proof-of-concept study has been done using ten simulated cases comparing a gastroenterology AI platform—GastroGPT, a specialty-specific AI platform vs other general AI models [OpenAIs GPT-4, Google's Bard, and Anthropic's Claude]. As expected, the scores in GastroGPT were higher. However, its is the ease of access which makes generic AI platforms accessible.[16]

In their paper, Cankurtaran *et al*. also looked at the reliability and utility scores of questions related to IBD using a 7-point Likert Scale. Questions used were those frequently searched online by patients and others directed towards healthcare professionals. The reliability scores of the answers for the professionals were significantly higher than those for the patients [$p = 0.032$]. The authors observed that the lowest scores in terms of reliability were in the treatment section.[17]

The major strengths of this study consist of the comprehensive set of questions collected over time from patients and worded by patients. The questions were given to us by two representatives of the patient association. This ensures that questions were asked in a way that patients would write or ask. To ensure an adequate assessment of ChatGPT replies, the answers were assessed independently by 14 different gastroenterologists with a special interest in IBD and who are well versed with the current literature and guidelines. To our knowledge, this is the first study to examine the accuracy and comprehensibility of ChatGPT's IBD-related medical knowledge.

A potential limitation of the study is a potential selection bias of a cohort of physicians limited to those in academic practice well versed in IBD and who were aware that these responses were generated by ChatGPT. Could their assessment have been stricter and the performance of ChatGPT underestimated? This is unlikely to have happened because for those answers which were marked low, there was always a plausible answer that was given, and the results were consistent. The median accuracy scores were slightly higher than mean scores [55.3%; Table 2]. This implies that even specialists may have slightly different thoughts in answering patients concerns. It would be interesting to note the replies

if these questions were answered by both IBD specialists and general gastroenterologists working in non-academic centres. Another limitation could be that the questions were given to us by patients from the patient group, and they may be of a higher education level.

In conclusion, this study provides evidence on the role of an AI-based system to provide accurate and comprehensive answers to real-world patient queries in the context of IBD. AI systems may serve as a useful adjunct tool for patients, in addition to the standard of care given in clinics. Such systems could also be a valuable resource for doctors working in remote areas. Apart from the refinement of the replies given by ChatGPT-3.5, more clarity is also required in how these answers were formulated, in terms of authorship, source, and date when the information was last updated. Furthermore, the replies need to more assertive, specific and with a firm recommendation and without re-referring to the caring specialist.

Funding

This research study did not receive any external funding. All aspects of the study, including data collection, analysis, and manuscript preparation, were conducted without financial support from any funding agency or organization.

Conflict of Interest

There was no conflict of interest declared by the authors pertaining to this study.

Author Contributions

MS, YF, and PE designed the study, recruited the participants, analysed the data, and wrote the draft and final version of the manuscript. HG, FF, MA, JT, NA, GF, MI, BV, FM, KK, JBKG, VAF, and SCZ contributed to the data, and analysed the draft and original article. All authors have read and approved the final version of the manuscript.

Data Availability

The datasets generated and analysed during the study are available by contacting the corresponding author: martina.sciberras.2@gov.mt.

Supplementary Data

Supplementary data are available online at *ECCO-JCC* online.

References

1. Sustersic M, Gauchet A, Foote A, Bosson J. How best to use and evaluate patient information leaflets given during a consultation: a systematic review of literature reviews. *Health Expect* 2016;**20**:531–42. doi:10.1111/hex.12487
2. EFCCA - EUROPEAN FEDERATION OF CROHN'S AND ULCERATIVE COLITIS ASSOCIATIONS. [cited October 25, 2023]. https://www.eu-patient.eu/Members/The-EPF-Members/Full-Membership/European-Federation-of-Crohns-and-Ulcerative-Colitis-Associations---EFCCA
3. Swire-Thompson B, Lazer D. Public health and online misinformation: challenges and recommendations. *Annu Rev Public Health* 2020;**41**:433–51. doi:10.1146/annurev-publhealth-040119-094127
4. Eysenbach G, Powell J, Kuss O, Sa E-R. Empirical studies assessing the quality of health information for consumers on the World Wide Web. *JAMA* 2002;**287**:2691–700. doi:10.1001/jama.287.20.2691
5. One in two EU citizens look for Health Information Online [Internet]. Eurostat; 2021 [cited October 25, 2023]. https://doi.org/10.2908/ISOC_CI_AC_I
6. Introducing chatgpt. [cited October 25, 2023]. https://openai.com/blog/chatgpt
7. Biswas SS. Role of chat GPT in public health. *Ann Biomed Eng* 2023;**51**:868–9. doi:10.1007/s10439-023-03172-7
8. Brown TB, Dhariwal P, Kaplan J, Subbiah M, Ryder N, Mann B. Language models are few-shot learners. 34th Conference on Neural Information Processing Systems. July 22, 2020. https://doi.org/10.48550/arXiv.2005.14165
9. Johnson D, Goodman R, Patrinely J, *et al*. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the chat-GPT model. *Res Sq* 2023:rs.3.rs-2566942. PMID: 36909565; PMCID: PMC10002821. doi:10.21203/rs.3.rs-2566942/v1
10. Lee T-C, Staller K, Botoman V, Pathipati MP, Varma S, Kuo B. CHATGPT answers common patient questions about colonoscopy. *Gastroenterology* 2023;**165**:509–11.e7. doi:10.1053/j.gastro.2023.04.033
11. Gilson A, Safranek CW, Huang T, *et al*. How does CHATGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ* 2023;**9**:e45312. doi:10.2196/45312
12. Cao JJ, Kwon DH, Ghaziani TT, *et al*. Accuracy of information provided by CHATGPT regarding liver cancer surveillance and diagnosis. *Am J Roentgenol* 2023;**221**:556–9. doi:10.2214/ajr.23.29493
13. Yeo YH, Samaan JS, Ng WH, *et al*. Assessing the performance of chatgpt in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol* 2023;**29**:721–32. doi:10.3350/cmh.2023.0089
14. Victora CG, Barros AJ, França GV, da Silva IC, Carvajal-Velez L, Amouzou A. The contribution of poor and rural populations to national trends in reproductive, maternal, newborn, and child health coverage: analyses of cross-sectional surveys from 64 countries. *Lancet Global Health* 2017;**5**:e402–7. doi:10.1016/s2214-109x[17]30077-3
15. Sallam M. CHATGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare* 2023;**11**:887. doi:10.3390/healthcare11060887
16. McCall B. GastroGPT outperforms general models in GI clinical tasks. 2023 [cited October 25, 2023]. https://www.medscape.com/viewarticle/997542?form=fpf
17. Cankurtaran RE, Polat YH, Aydemir NG, Umay E, Yurekli OT. Reliability and usefulness of ChatGPT for inflammatory bowel diseases: an analysis for patients and healthcare professionals. *Cureus* 2023;**15**:e46736. doi:10.7759/cureus.46736

Can we simplify the journey in UC?

JYSELECA is a once-daily oral treatment* that provides rapid** and long-term† efficacy up to ~4 years[1–3]

Helping patients return to their normal lives[4††]

Discover more

Full Prescribing information. Report an adverse event.

* Recommended dose for induction and maintenance is 200 mg once daily.[1] JYSELECA is not recommended in patients aged 75 years and older as there is no data in this population; in patients aged 65 years and over the recommended dose is 200 mg once daily for induction treatment and 100 mg daily for maintenance treatment.[1]
** Data from a *post-hoc* analysis of diary data from the double-blind, randomised, placebo-controlled 58-week SELECTION trial. Achievement of stool frequency subscore of ≤1 by Day 3 in biologic-naïve patients, and rectal bleeding subscore of 0 by Day 5 in biologic-experienced patients.[2]
† Interim analysis of SELECTIONLTE assessing the efficacy and safety of open-label JYSELECA 200 mg through LTE Week 144 in completers and LTE Week 192 in non-responders, respectively, representing a total of 3.9 years of treatment each (completers: 58 + 144 weeks; non-responders 10 + 192 weeks).[3]
†† Determined in a *post-hoc* exploratory analysis of the SELECTION trial assessing HRQoL and the comprehensive disease control multi-component endpoint, which comprises both clinical and QoL outcomes, in individuals receiving JYSELECA (n=786).[4] Each patient has their own definition of normal life.

▼ This medicine is subject to additional monitoring.

HRQoL, Health-related quality of life; LTE, Long term extension; QoL, Quality of life; UC, Ulcerative colitis.

1. JYSELECA Summary of Product Characteristics, January 2024.
2. Danese S, et al. Am J Gastroenterol 2023;118(1):138–147.
3. Feagan BG, et al. ECCO 2023; #OP35.
4. Schreiber S, et al. J Crohns Colitis 2023;17(6):863–875.

Galápagos
Pioneering for patients

GL-UC-FIL-202312-00012 | January 2024

Jyseleca® ▼
filgotinib