89525_Auto_Edited-check.docx

WORD COUNT	3951	TIME SUBMITTED	27-DEC-2023 06:55PM
		PAPER ID	105769181

Name of Journal: *World Journal of Gastroenterology*

Manuscript NO: 89525

Manuscript Type: EVIDENCE REVIEW

May ChatGPT be a tool producing medical information for common inflammatory bowel disease patients' questions? An evidence-controlled analysis

Gravina AG *et al*. ChatGPT for IBD patients

Abstract

Artificial intelligence is increasingly entering everyday healthcare. Large language model (LLM) systems such as Chat Generative Pre-trained Transformer (ChatGPT) have become potentially accessible to everyone, including patients with inflammatory bowel diseases (IBD). However, significant ethical issues and pitfalls exist in innovative LLM tools. The hype generated by such systems may lead to unweighted patient trust in these systems. Therefore, it is necessary to understand whether LLMs (trendy ones, such as ChatGPT) can produce plausible medical information (MI) for patients. This review examined ChatGPT's potential to provide MI regarding questions commonly addressed by patients with IBD to their gastroenterologists. From the review of the outputs provided by ChatGPT, this tool showed some attractive potential while having significant limitations in updating and detailing information and providing inaccurate information in some cases. Further studies and refinement of the ChatGPT, possibly aligning the outputs with the leading medical evidence provided by reliable databases, are needed.

Key Words: Crohn's disease; Ulcerative colitis; Inflammatory bowel disease; Chat Generative Pre-trained Transformer; Large language model; Artificial intelligence

Gravina AG, Pellegrino R, Cipullo M, Palladino G, Imperio G, Ventura A, Auletta S, Ciamarra P, Federico A. May ChatGPT be a tool producing medical information for common inflammatory bowel disease patients' questions? An evidence-controlled analysis. *World J Gastroenterol* 2023; In press

Core Tip: Patients with inflammatory bowel disease (IBD) increasingly access information resources online to receive information about disease management. Emerging artificial intelligence (AI) systems such as Chat Generative Pre-trained Transformer (ChatGPT) are taking hold in the daily reality of many patients with IBD. Through them, patients can potentially understand these systems as reliable or

substitutes for medical consultation, especially for issues about reluctantly talking to their gastroenterologist. This review, examining ChatGPT's outputs to common questions from patients with IBD, outlined how, while this AI system can provide some medical information, there are some limitations related to poor updating and the risk of inaccuracies that push for its cautious use.

INTRODUCTION

The Chat Generative Pre-trained Transformer (ChatGPT) (www.chat.openai.com) is an artificial intelligence (AI)-based conversational large language model (LLM) chatbot system developed by OpenAI (San Francisco, CA, United States) and released in November 2022[1]. ChatGPT sparked a vigorous debate in the scientific community regarding the application of AI in the scientific literature (*e.g.*, in the writing of scientific articles) by bringing the spotlight to bear on the scientific reliability and accuracy that such a system could offer[2,3].

ChatGPT has also been called upon as a possible bot to answer patients' questions regarding their diseases, offering, in some cases, the potential for this purpose[4,5]. In gastroenterology, the possible application of ChatGPT is still highly pioneering, little explored, and far from being codified. There has been an interest in ChatGPT in the gastroenterology community, especially in the possibility of being able to answer clinical questions posed by patients and research questions. Concerning the latter, Lahat *et al*[6], for example, expressed some potential of LLMs in the genesis of research questions, although there is a great need to improve their novelty. Yeo *et al*[7], on the other hand, showed promising results of ChatGPT in answering clinical questions about liver cirrhosis. Similar results have recently been reported regarding colonoscopy-related medical information (MI)[8].

In inflammatory bowel disease (IBD), medical communication with the patient is crucial, as these diseases affect the patient to three hundred and sixty degrees by directly affecting their quality of life. IBDs are chronic, relapsing-remitting diseases with a particularly complex and multifactorial pathogenesis, mainly including Crohn's

disease (CD) and ulcerative colitis (UC). Therefore, patients must undergo periodic medical check-ups, diagnostic tests, and courses of treatment, often for a lifetime[9,10].

Consequently, in today's context of widely available technology, patients often access information technology resources to obtain information for managing their IBD. The Internet is a prime tool for this purpose because it offers the patient a considerable window of resources, including social media[11,12] and ChatGPT. Patients often consult these resources independently to conduct targeted research for their concerns, but the physician is often integrated into this process through telehealth tools[13]. This analysis aimed to review the scientific validity of AI-generated outputs provided by ChatGPT regarding the genesis of MI regarding ten questions raised by patients with IBD.

GENERAL CONSIDERATIONS

ChatGPT is a promising tool with some baseline limitations to consider at the outset but with some promising advantages

ChatGPT is based on a natural language processing model developed by OpenAI, which allows the user to use it for various operations such as chatbots, dialogue systems, text formation, and question-answering[14]. Different versions of ChatGPT (*i.e.*, GPT 1, 2, 3, and 4) have been developed over time, and it has been observed that it has grown from 117 million programming parameters (in GPT 1) to 300 billion parameters in GPT-3 with exponential improvement in various tasks (*i.e.*, fine-tuning datasets, fine-tuning tasks, language understanding, text generation, and sentiment analysis)[14,15]. ChatGPT is based on a "training model" for reinforcement learning[14].

However, because ChatGPT is available to everyone, special care must be taken when such a platform is used by both healthcare professionals and patients to produce MI. ChatGPT has several limitations that have already been postulated. These include the lack of contextual understanding, the lack of common sense, the dependence of information on the need to provide the system with large amounts of data, and the lack of interpretability[14].

Ultimately, ChatGPT has limited knowledge because its operation (based on data-driven training processes) depends on the data on which it has been trained; thus, its merits do not include constant updating[16]. This is a significant limitation when approaching MI, as medical knowledge is highly changeable and is significantly affected by daily updates.

Because of these limitations, some authors have emphasised developing advanced LLMs to benefit patients with error-free MI[17]. In some medical contexts, however, ChatGPT has proven in early studies to perform better than other mainstream search engines (*e.g.*, Google search)[18].

In contrast, the ChatGPT has several advantages. These include the availability of an always-on service with no downtime, a fast system with some efficiency, the ability to speak several languages (expanding the user base to include non-English-speaking people), and lastly, it is not operator-dependent as an AI[16].

ChatGPT's ability to produce human-like language led to the theorising that LLMs could represent an apparent revolution in healthcare[19]. The same applies to human-like problem-solving skills[20]. However, in a healthcare setting, especially a chronic one such as IBD, one of the biggest challenges is dealing with the empathy skills of the ChatGPT[5]. The empathy physicians can create with patients with IBD is crucial to the physician-patient relationship[21-24]. Patients with IBD are known to suffer from a high prevalence of anxiety-depressive disorders even in remission[25,26]. Figure 1 summarises some general advantages and disadvantages of using ChatGPT-generated AI.

Selection of ChatGPT inputs for evidence review in the scientific literature and major IBD guidelines

A group of IBD-expert physicians retrieved a list of ten questions most frequently asked by patients with IBD (related to their IBD management) in their current clinical care practice. The ten with the highest frequency (Q1-10) were collected from the total number of questions. The people selecting inputs were not restricted in their choice of questions or given specific filters to adopt, but the only guideline provided was to select

questions that patients asked most frequently in their current clinical practice. This mechanism was intended to sample real questions asked by IBD patients and not hypothesized/thought by physicians to avoid biased questions generated by a person with IBD scientific expertise. These questions were then input on ChatGPT on three different days (18th, 19th, and 20th August 2023), and each output generated by the chatbot was categorised as O1, O2, and O3, respectively. All research staff belonged to the Hepatogastroenterology Division of the University of Campania Luigi Vanvitelli, a regional Italian referral hospital for the management of IBD. All physicians involved in the study regularly contacted patients with IBD in their daily clinical practice.

The same research team evaluated the AI-generated responses by ChatGPT for each question by objectively comparing them with the available evidence. The ten questions with the highest frequency provided by all gastroenterologists in the study are listed in Table 1, and the ChatGPT outputs are listed in Table 2.

EVIDENCE REVIEW: WHAT ARE THE RESPONSES OF CHATGPT?

Q1 - ChatGPT provides correct information on the existence or non-existence of definitive therapy for IBD, albeit with a paucity of detail

The first input (Q1) concerned the potential existence of a definitive therapy for IBD. Q1 ChatGPT outputs (O1-3) correctly defined IBD, expressing their chronicity, the main phenotypes (*i.e.,* CD and UC), and the target of their inflammatory action (*i.e.,* the gastrointestinal tract)[27]. The outputs correctly expressed the absence of definitive therapy for IBD, and Q1 O1 outlined the macro categories of treatments currently available for IBD (*i.e.,* medical and surgical treatments)[28-31].

In addition, the goals toward which specific IBD therapy should strive provided by the outputs (*i.e.,* induction/maintenance of remission, prevention of complications, and improvement of quality of life) are the focus of the European Crohn's and Colitis Organisation guidelines[28-31], as of the current consensus on Selecting Therapeutic Targets in IBD[32]. The final aspect of Q1 O1 is how optimal nutrition also affects the therapeutic management of IBD[33]. Q1 O2,3 did not address this aspect. In all outputs,

ChatGPT set out the need to keep up to date with the pace of research and consult a health professional out of necessity.

Approaching this subject in the case of UC is particularly difficult. Surgical treatment of UC with definitive techniques (*e.g.*, restorative proctocolectomy without ileostomy) does not always guarantee the absence of short- and long-term complications[31]. For example, packing the ileal pouch can lead to the emergence of acute and chronic forms of pouchitis[34]. In contrast, in CD, surgery does not exclude the reactivation of the disease at the perianastomotic site or its emergence at other gastrointestinal sites[35].

Q1 O1-3, therefore, failed to make a clear distinction between CD and UC in terms of the power to control the inflammatory burden of IBD by not expressing the different possibilities that surgery can offer between CD and UC. In other words, in conclusion, Q1 O1-3 have not been able to fully filter the nuances that exist between healing and cure, adapt these concepts to the IBD phenotype, and grade the therapeutic approach to the curative degree it can provide (especially in conditions such as UC where surgery drastically adjusts the course of the disease and its natural history by healing the underlying disease).

Q2 - ChatGPT provides dietary principles for IBD patients while not making explicit the limited amount of evidence available for the role of nutrition in many aspects of IBD management

Q2 focused on the kind of nutrition the IBD patient should follow to correctly manage his or her disease. In the case of Q2, ChatGPT provided, in all outputs, a list of dietary advice; however, considering it as a prerequisite for such advice, the need for patient to seek professional advice. The proposed dietary advice included a low-residue diet, low fermentable oligosaccharides, disaccharides, monosaccharides, and polyols (FODMAP) diet, anti-inflammatory foods, lean protein, good fats, hydration, avoidance of trigger foods, small and frequent meals, probiotics, monitoring fiber intake, dairy alternatives, and supplements.

Q2, O2, and O3 also advised personalising the diet to identify trigger foods (*i.e.*, "listen to your body") and working with professionals. O2 also advised avoiding dietary changes too quickly. Comprehensive nutrition analysis in IBD is particularly complex because no specific diet can be specifically recommended to induce remission in patients with active disease, as stated by the European Society for Clinical Nutrition and Metabolism guidelines[33]. The outputs' premise differed in that O1 defined the provided list of nutritional recommendations as "commonly recommended for managing IBD" while still advising to counsel with a professional. In contrast, Q2 O2-3 were more reluctant to define specific recommendations. However, the list of dietary advice provided seems to appear as "generally deemed" dietary advice. The evidence underlying such advice from the perspective of the "safety/efficacy" profile is, as written, particularly poor[33].

In Q2, as in Q1, ChatGPT failed to detail some aspects of IBD nutrition. For example, IBD nutrition can be resent from watersheds, especially in patients who are already undergoing surgery. In addition, although a low-fiber diet is recommended in Q2 O1-O3, such a regimen is not always valid for all patients with IBD, but it can be considered in cases such as patients with CD with a stricturing phenotype[36]. In addition, it is still complex to isolate which specific dietary components (*e.g.*, cereals, sugar, fat, protein, and dietary fiber) may be associated with relapse or worsening of pre-existing clinical manifestations in IBD[36].

Indeed, stepping outside the realm of specific guideline recommendations, the low-FODMAP diet, while giving good results on symptomatology control in several IBD-focused studies, has not yet been firmly proven to reduce gut inflammation and, indeed, in some settings, has reduced several favourable bacterial species (*i.e., F. prausnitzii*, *C. cluster IV*)[37].

Finally, ChatGPT correctly instructed the patients to hear from their professional caregivers before taking supplements or probiotics. This is because of the often disproportionate and misguided use of supplements in patients with IBD[38] and the poorly defined evidence on the benefits of consuming specific probiotics[39-41].

Q3 - ChatGPT provides fair indications for performing or repeating endoscopic examinations in patients with IBD and does not provide a specific frequency of repeat examinations

One of the most severe issues for IBD patients is undoubtedly the need to undergo repeated endoscopic examinations to manage their disease, especially when performed under conditions of clinical remission (*e.g.,* colorectal cancer surveillance)[42]. ChatGPT has adopted a particularly weighted approach to answer this question. In Q3, O1 differed significantly from O2,3. O1 deferred the discussion to the gastroenterology specialist, expressing an extreme variety of factors determining the frequency of endoscopic examinations. O2,3, on the other hand, also explicitly listed several cases in which colonoscopy can be performed or repeated to manage IBD. Such cases include initial diagnosis, disease activity monitoring, flare-up, surveillance (*i.e.,* long-standing UC), and after surgery. All possible indications of performing/repeating lower gastrointestinal endoscopic examinations proposed by ChatGPT are scientifically supportable[43,44]. Even in Q3, in each output, ChatGPT expressed the need for the patient to refer to their specialist and did not launch improper definitions of colonoscopy repeat frequency (since Q3 was a direct question about the number of times to repeat colonoscopy).

Q4 – ChatGPT's ability to respond to patient demands for therapeutic changes: Enemas in UC-possible risks of deterrence and inaccuracy

Traditional therapies still play a crucial role in UC management. Central to the latter is 5-aminosalicylic acid (5-ASA), which can be administered in oral formulations (with different delivery techniques based on Eudragit or MMX) and topically. Topical formulations allow the direct attack of proctitis-type (*i.e.,* E1) or distal UC (*i.e.,* E2) forms of UC according to the Montreal classification[29].

Recent meta-analytic evidence showed that combination therapy with topical and oral 5-ASA had the highest performance for induction of clinical remission (P-score

0.91), prevention of recurrence (P-score 0.91), and induction of endoscopic remission (P-score 0.9) while showing an optimal safety profile[45]. Nevertheless, 5-ASA also has the ability (with a minimum dosage of 1.2 g daily) to possess a chemopreventive effect against colorectal cancer (odds ratio = 0.46)[46]. These premises make the choice of modifying and/or removing topical therapy in UC difficult, considering the short- and long-term benefits it can provide in both the inductive and maintenance phases of remission. Conversely, the patient does not always easily tolerate topical therapy, which can provide discomfort[47-50].

In Q4, the input asked ChatGPT about the possibility of not performing topical therapy for UC. Q4 O1 correctly explained that the use of enemas is part of the therapeutic possibilities of UC and that it depends on disease activity and declined the choice of the removal of enemas to the healthcare professional[29,36]. Q4 O2,3 also expressed the alternatives to enemas that the patient can discuss with their specialist (*i.e.,* oral therapy, suppositories, lifestyle modifications, surgery) and in the case of O3 biological therapy or topical foams/creams.

Q2 O2,3, however, in listing alternatives to enemas, ignored the aspects already written in the introduction to this paragraph, namely the evidence that topical 5-ASA therapy (especially when combined with oral therapy) is still highly relevant. Q2 O1-3 also ignored how enemas were administered and the importance of correctly delivering them. It would have been helpful (while advising referral to the physician) to provide a list of tips on how to perform enemas properly, the possibility of adjusting the volume of enemas according to individual tolerability, and advice on how to avoid local soreness or early evacuation of the enema. In other words, Q4 O1-3 may have too high a deterrent load toward enemas, prompting the patient not to do so.

In addition, Q4 O2 also provided inaccuracy in terms of oral therapy (*i.e.,* "These drugs can help reduce inflammation and manage symptoms without the need for enemas"). This assertion contradicts the available evidence[29,45]. In addition, in the same output, ChatGPT asserts "These medications work similarly to enemas but are inserted into the rectum", referring to suppositories. This is further evidence of inaccuracy[36]. Q4

O2-3 cannot be provided to the patient and should be subjected to a physician's preliminary filter.

Q5 - ChatGPT responds to the IBD patient's desire for pregnancy: A good brochure to prepare for such a patient's decision

IBD affecting female patients of reproductive age can collide with pregnancies that such patients may develop. However, for this reason, there are concerns that patients manifest as many IBD-related as not[51-53].

In this sense, Q5 obtained outputs expressing the need to work side-by-side with specialists in the field (*e.g.,* obstetricians/gynaecologists and gastroenterologists). The outputs also expressed some factors to consider when planning a pregnancy affected by IBD. ChatGPT correctly listed some essential factors in the pregnancy pathway in the context of IBD. In detail, they include pre-conception counselling, the preferability of having a pregnancy with well-controlled disease activity, and the need to tailor IBD therapy to safety for the foetus in prenatal care[54].

Another positive aspect is that ChatGPT, in all outputs related to Q5, expressly specified that the mode of delivery (vaginal or caesarean) is affected by IBD. It is well known that the non-vaginal mode of delivery should be preferred in patients with active perinatal disease, after restorative proctocolectomy, or with rectovaginal fistula[54].

Q6 - The patient demand for non-parenteral formulations of biologic drugs highlights ChatGPT's lack of capacity for constant updating: No mention of new orally administered molecules

In Q6, the question was asked about the possibility of receiving biologic drug administration for IBD treatment in the oral *vs* the parenteral form (*i.e.,* intravenous or subcutaneous). ChatGPT responses (updated until September 2021) were far from the current reality, where oral formulations for IBD, particularly approved for UC, are also

available in guideline-recommended indications (*e.g.,* tofacitinib, filgotinib, ozanimod, upadacitinib)[29,55,56].

Q7 - A major concern for the patient with IBD: The risk of colorectal cancer: ChatGPT answers direct the patient to individual risk stratification and the physician-guided need for regular surveillance

Q7 receives slightly different outputs. O1 was the most in-depth and listed some factors that influence the increased risk of colorectal cancer in IBD (especially in UC), such as, by way of example, duration of disease and severity of inflammation and family history, as well as how it is necessary to undergo regular surveillance colonoscopy and IBD therapy because of the action of some therapies for IBD in colorectal cancer management. Q7 O2 and O3 expounded these concepts more succinctly than O1 did. ChatGPT answers are scientifically plausible because, to date, European guidelines recommend risk stratification into three risk categories (*i.e.,* lower, intermediate, and high) based on several parameters (*e.g.,* extent of disease, familiarity, presence of stenosis/dysplasia, and extra-intestinal manifestations such as primary sclerosing cholangitis) and, based on these, determine the frequency with which patients with IBD should undergo surveillance colonoscopy[57].

Q8: ChatGPT sufficiently weighs communication about the risk of cancer or infection related to biological therapy in IBDs

The prevalence of colorectal cancer or colic dysplasia in IBD is far from insignificant and is responsible for 10%-15% of deaths in patients with IBD[58]. The risk is especially high in UC patients with extensive disease after 8-10 years of diagnosis[57]. In addition, undergoing some biological therapies may increase the risk of infectious events[59].

Regarding the risk of biologics-related cancer, ChatGPT outputs were cautious, expressing that risk was generally low, stigmatizing the need for this risk to be weighed individually for the risk-benefit ratio by the health care specialist[60-63]. Regarding

infectious risk, outputs have correctly expressed how it may be increased during biological therapy[59].

Q9 – Patient's fear of heritability of their IBD to their children: ChatGPT clarifies that genetics is only part of the whole ("a piece of the puzzle") and that IBD is not a traditional genetic disorder

Q9 O1-3 correctly included genetics as one of many components that can determine the pathogenesis of IBD[27] and did not guarantee the patient a heritability of IBD to offspring as a canonical genetically transmitted disease[64]. In addition, ChatGPT, in its outputs, also added how family history, a well-known risk factor for IBD, plays a role in determining the heritability of IBD[65,66]. As in all outputs related to all questions, in the case of Q9, ChatGPT referred the patient to a specialist in the field. In this specific question, it avowed the role of the geneticist.

Q10 - Squeezing ChatGPT's specialized expertise: The patient asks about the need for biologic therapy after the first CD surgery

The last question (*i.e.,* Q10) exposed the possibility of avoiding the biological therapy set for prophylaxis of postoperative recurrence in CD. Q10 O1 did not comment and referred the discussion to the physician. Q10 O2,3, on the other hand, correctly exhibited some of the factors based on which one may decide to avoid biological therapy after initial CD surgery (*e.g.,* extent of surgery, risks and benefits, patient preference)[35].

CONCLUSION

To the best of our knowledge, this review was the first of its kind to weigh the ability of ChatGPT, an LLM system, to provide AI-generated reliable MI. A comparison of the outputs provided by ChatGPT showed that such AI-generated MI shows some scientific reliability. However, our analysis shows that this judgment is not always valid and

expected in that some outputs (*e.g.*, those related to Q4 or Q6) were not scientifically correct or poorly updated (*i.e.*, until September 2021).

The use of LLM and AI-generated MI is, at present, the subject of intense controversy in the scientific community in a dichotomy between revolution in medical education, research and patient communication until concerns related to a new potential infodemic[1-3,67-69]. This review expresses the need for further refinements of ChatGPT before it can be implemented as a complementary response mechanism to patient concerns.

In addition, it should be made sure that ChatGPT is trained on papers provided by databases commonly considered valid in the scientific community (*e.g.*, Scopus, Web of Science, MEDLINE). Another essential pitfall is the lack of updating that ChatGPT detects in almost all its outputs referred to until September 2021 (Figure 2). In addition, ChatGPT did not provide evidence levels for the sources employed to generate the outputs, thus removing the reader's ability to discern different degrees of quality for the same. Finally, it is also necessary to explore the capability of ChatGPT in so many other aspects related to IBD not already covered in this review (the latter, moreover, schematized in Figure 3).

89525_Auto_Edited-check.docx

ORIGINALITY REPORT

0%

SIMILARITY INDEX

PRIMARY SOURCES

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	< 15 WORDS
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	< 15 WORDS