

Evaluating ChatGPT-5o as a Clinical Decision-Support Tool in Inflammatory Bowel Disease: A Pilot Study of Guideline Adherence and Clinical Agreement

Daniel Aillaud-De-Uriarte^{a, b, i} , Alejandro Cendejas-Higuera^c , Eric Acosta-Marquez^d ,
Luis A. Hernandez-Flores^e , Monica Reyes-Bastidas^{f, g} ,
Haide Manzano-Cortes^h 

Abstract

Background: Inflammatory bowel disease (IBD) presents complex management challenges. While care is guided by expertise and guidelines, artificial intelligence (AI) is being explored as an adjunct. This study evaluates ChatGPT-5o's ability to provide IBD recommendations by comparing its outputs with real-world decisions and European Crohn's and Colitis Organisation (ECCO) guidelines.

Methods: We performed a retrospective analysis of 19 anonymized IBD cases spanning initial and complicated disease. ChatGPT-5o generated management recommendations, which were compared with clinician treatments and ECCO guidelines across seven therapeutic domains (5-aminosalicylic acid (5-ASA), steroids, antibiotics, thiopurines, anti-tumor necrosis factor (TNF), anti-integrins, anti-interleukin-23 (IL-23)) plus diagnostic workup, symptom management, surgical consultation, and monitoring. Agreement was quantified using Cohen's Kappa.

Results: ChatGPT-5o showed perfect agreement ($\kappa = 1.000$) with providers and/or guidelines for antibiotics, diagnostic workup, symptom management, surgical consultation, monitoring, and anti-IL-23. Substantial agreement ($\kappa \approx 0.6 - 0.8$) was observed for 5-ASA and steroids. Moderate to fair agreement ($\kappa \approx 0.3 - 0.5$) occurred for anti-

TNF and anti-integrins, reflecting variability in complex scenarios. Thiopurines demonstrated the lowest concordance, with none-to-slight agreement in human-AI comparisons but higher alignment of ChatGPT-5o with ECCO, suggesting evolving practice patterns and safety considerations.

Conclusions: ChatGPT-5o closely aligns with clinicians and ECCO guidelines in multiple standardized domains, supporting its potential as a decision-support tool to enhance guideline adherence and broaden access to IBD expertise. Variability in biologic selection and thiopurine use underscores the need for expert oversight and patient-specific judgment. Prospective studies should assess longitudinal outcomes and integration strategies to ensure safe, patient-centered deployment.

Keywords: Inflammatory bowel disease; Crohn's disease; Ulcerative colitis; Artificial intelligence; Clinical decision support; Machine learning; Large language models

Introduction

Inflammatory bowel disease (IBD), which includes Crohn's disease (CD) and ulcerative colitis (UC), is a chronic relapsing inflammatory condition of the gastrointestinal tract that has become a major global health concern. Its incidence and prevalence have increased dramatically in recent decades, especially in industrialized countries, placing a growing burden on healthcare systems and driving efforts to better understand its etiology, diagnosis, and management [1]. Although the exact cause of IBD remains unknown, it is widely accepted that the disease arises from a complex interaction between genetic susceptibility, environmental triggers, gut microbiome disruption, and immune dysregulation [2]. Interestingly, IBD rates are rising even in previously low-prevalence areas such as Latin America and Asia, underscoring the influence of urbanization, diet, and other environmental factors [3].

In North America and Europe, incidence rates can range from 10 to over 40 cases per 100,000 individuals, while lower rates are still reported in South America and Asia [3]. But this is changing quickly. Risk factors like smoking, diet, stress, and

Manuscript submitted October 27, 2025, accepted November 20, 2025
Published online December 15, 2025

^aCenter for Bioethics, Harvard Medical School, Boston, MA, USA

^bSchool of Public and Population Health, Institute for Bioethics and Health Humanities, The University of Texas Medical Branch (UTMB), Galveston, TX, USA

^cSchool of Medicine, Universidad de Monterrey, Monterrey, Mexico

^dSchool of Medicine, Universidad Anahuac Puebla, Puebla, Mexico

^eDepartment of Surgery, Houston Methodist Hospital, Houston, TX, USA

^fDepartment of Gastroenterology, Hospital Angeles Culiacan, Culiacan, Mexico

^gDepartment of Gastroenterology, Hospital General Regional No.1 Instituto Mexicano del Seguro Social (IMSS), Culiacan, Mexico

^hDepartment of Gastroenterology, Hospital Angeles Puebla, Mexico

ⁱCorresponding Author: Daniel Aillaud-De-Uriarte, Center for Bioethics, Harvard Medical School, Boston, MA, USA.

Email: daniel_aillauddeuriarte@hms.harvard.edu

doi: <https://doi.org/10.14740/aicm12>

nonsteroidal anti-inflammatory drug (NSAID) use have all been studied - smoking, for example, worsens Crohn's disease but may offer a protective effect in ulcerative colitis [4]. These trends highlight the need for early diagnosis, multifaceted prevention strategies, and treatment plans that are sensitive to both population-level data and individual circumstances.

IBD management has become increasingly complex. While the introduction of biologics, small molecules, and combination therapy has improved outcomes for many patients, therapeutic responses vary, and resistance is not uncommon [5]. Standardized guidelines developed by expert societies like the European Crohn's and Colitis Organisation (ECCO) help clinicians navigate this complexity, but in practice, care often deviates from these recommendations due to clinical judgment, patient-specific considerations, or access-related issues [6, 7].

At the same time, artificial intelligence (AI) has begun to play a growing role in medicine. In gastroenterology, AI has been applied to polyp detection, endoscopy image interpretation, and risk prediction tools [8]. For IBD specifically, machine learning models have shown promising results in predicting outcomes like hospitalization, need for surgery, or steroid dependence [9]. However, these tools have typically relied on structured electronic medical record (EMR) data, rather than generating full-text clinical recommendations.

This is where large language models (LLMs) like ChatGPT come in. These models can process natural language and generate treatment plans based on user input. The idea that such tools could assist in clinical decision-making - especially in settings with limited access to IBD expertise - is both exciting and controversial [10, 11]. AI recommendations may lack nuance, overlook patient context, or reinforce existing biases, raising important questions about their place in real-world care [12].

This pilot study aims to evaluate how ChatGPT-5o performs in providing IBD management advice. By comparing its responses to ECCO guidelines and clinician recommendations for a set of clinical vignettes, we assess whether generative AI can be trusted as a decision-support tool in IBD. We focus on key therapeutic areas such as 5-aminosalicylic acids (5-ASAs), corticosteroids, immunomodulators, biologics, and antibiotics, and examine the degree of concordance between ChatGPT-5o's outputs, expert consensus, and clinical judgment. In doing so, we hope to better understand not only what generative AI gets right, but also where it falls short - and what that means for its potential integration into gastroenterology practice.

Materials and Methods

This was a retrospective comparative study evaluating the agreement between ChatGPT-5o (OpenAI) and practicing gastroenterologists in the management of IBD, including both CD and UC. Clinical case scenarios were derived from EMR and included both initial presentations and complex disease phases. An earlier version of this analysis was presented at ECCO 2025 using ChatGPT-4o. For the present manuscript, all 19 cases were re-run and verified with ChatGPT-5o, the

updated model available at the time of study preparation. We selected ChatGPT-5o because it represents one of the most widely recognized and publicly accessible LLMs currently available, with prior versions already used in clinical and academic settings. Its broad visibility and integration into health-care research make it a relevant reference point for assessing real-world clinical decision-support potential. Moreover, using the most updated, stable release available at the time reduced cross-model variability and ensured consistency with our earlier pilot study using ChatGPT-4o.

This study was designed as a pilot exploratory analysis, and the 19 cases included were selected to capture a wide range of real-world clinical scenarios, from initial presentations to complex disease states. Although the sample size is modest, it was sufficient for evaluating agreement patterns across multiple therapeutic domains using Cohen's Kappa statistics, which is appropriate for small datasets. Future studies will expand the number of cases and include comparisons across multiple LLMs to validate and generalize these findings.

A total of 19 anonymized IBD cases were selected from patient records. The cohort included newly diagnosed patients, individuals in remission, and those with complications such as strictures, fistulas, or extraintestinal manifestations. Cases were purposively selected to represent a wide spectrum of real-world IBD presentations, including both CD and UC, initial and complicated disease courses, and varied severities. The goal was to capture the diversity of treatment decisions encountered in clinical practice rather than to analyze a strictly consecutive or randomized series. This approach aligns with the exploratory, pilot nature of the study. No patient-identifiable information was included, and ethical oversight was observed in accordance with local institutional review practices.

ChatGPT-5o was prompted using standardized case summaries to assess, diagnose, and recommend medical or surgical treatment in line with contemporary clinical standards. Prompts were phrased to simulate a clinical consultation, asking for initial impressions, diagnostic workups, and treatment plans aligned with current ECCO guidelines. The initial instructions presented to ChatGPT-5o were standardized: "We want to analyze your response and knowledge of medical treatment of IBD patients. We may provide you with the information available, including personal history, family history, medications, and images. If you need any extra information, please request it. Suggest medical or surgical treatment if required, doses, or change of medications, depending on your consideration." The model had no access to external resources during this process. The complete prompt text, model identifier, and additional technical details are available in Supplementary Material 1 (aicm.elmerpub.com).

Evaluations were conducted by two independent gastroenterologists, both members of the ECCO and the American Gastroenterological Association (AGA) and certified by the Mexican Association of Gastroenterology. They independently evaluated each case and rated the appropriateness of management decisions made by both human clinicians and ChatGPT-5o, based on the ECCO guidelines.

Agreement with ECCO guidelines was assessed across seven therapeutic categories: 5-ASA, corticosteroids, antibiot-

Table 1. Interpretation of Agreement Levels

Agreement level	Interpretation
Perfect (Kappa = 1.000)	Complete alignment, no variation between recommendations.
Substantial (Kappa $\approx$ 0.6 - 0.8)	High agreement with some minor differences, reflecting consistent recommendations.
Moderate to fair (Kappa $\approx$ 0.3 - 0.5)	Notable variability, indicating different interpretations or thresholds in decision-making.
Non-significant	Agreement level is not statistically meaningful; recommendations may vary substantially.

ics, thiopurines, anti-tumor necrosis factor (TNF) agents, anti-integrins, and anti-interleukin-23 (IL-23) agents. In addition, agreement was analyzed for decisions on diagnostic workup, symptom management, surgical consultations, and monitoring strategies. These domains were selected based on their central role in both the induction and maintenance phases of IBD treatment and were representative of guideline-directed care in real-world settings.

To quantify agreement between the different sources of recommendations (human clinicians, ChatGPT-5o, and ECCO guidelines), Cohen's Kappa coefficient was used for pairwise comparisons. Kappa values were interpreted using a modified Landis and Koch scale (Table 1). A P-value < 0.05 was considered statistically significant. All statistical analyses were conducted using IBM SPSS Statistics for Windows, Version 28.0.

All cases were anonymized and selected under ethical oversight consistent with institutional guidelines. No identifiable patient data were used. This study was determined as a not human research study by Harvard Human Research Protection Program; therefore, Institutional Review Board (IRB) review was not required. This study was conducted in accordance with institutional ethical standards and the principles of the Helsinki Declaration. No animal studies were conducted.

Results

Across the 19 IBD cases analyzed, agreement between Chat-

GPT-5o, human clinicians, and the ECCO guidelines varied by treatment domain. While certain areas such as diagnostic workup, symptom management, and anti-IL-23 therapy demonstrated perfect concordance, others, particularly involving thiopurines, anti-TNF agents, and anti-integrins, showed notable variability.

Cohen's Kappa coefficient was used to assess agreement across three pairwise comparisons: human clinicians vs. ECCO guidelines, human clinicians vs. ChatGPT-5o, and ChatGPT-5o vs. ECCO guidelines. A detailed breakdown of Kappa values across domains is presented in Table 2.

Perfect agreement ($\kappa = 1.000$)

Perfect agreement was observed in the domains of antibiotic use, diagnostic workup, symptom management, surgical consultation, continuous monitoring, and anti-IL-23 therapy. For these categories, ChatGPT-5o and human recommendations were identical to the guidelines or each other, showing complete alignment in decision-making.

Substantial agreement ($\kappa \approx 0.6 - 0.8$)

Substantial agreement was seen in the use of 5-ASA and corticosteroids, with Kappa values ranging from 0.578 to 0.771. Recommendations by ChatGPT-5o and/or humans were most-

Table 2. Summary of Cohen's Kappa Agreement for Each Category

Category	Human vs. guidelines Kappa	Human vs. ChatGPT Kappa	ChatGPT vs. guidelines Kappa	Agreement level	Significance
5-ASA	0.689	0.689	0.604	Substantial	Yes
Steroids	0.406	0.578	0.771	Moderate to substantial	Yes
Anti-TNF	0.289	0.345	0.685	Fair to substantial	Yes
Anti-integrins	0.486	0.336	0.771	Fair to substantial	Yes
Anti-IL-23	1.000	1.000	1.000	Perfect	Yes
Thiopurines	0.441	0.313	0.771	Fair to substantial	No
Antibiotics	0.642	1.000	1.000	Substantial to perfect	Yes
Diagnostic workup	1.000	1.000	1.000	Perfect	Yes
Symptom management	-	1.000	1.000	Perfect	Yes
Surgical consult	-	1.000	1.000	Perfect	Yes
Continuous monitoring	-	1.000	1.000	Perfect	Yes

5-ASA: 5-aminosalicylic acid; IL: interleukin; TNF: tumor necrosis factor.

Table 3. Explanation of Cohen's Kappa Agreement for Each Category

Category	Human vs. guidelines Kappa	Human vs. guidelines P-value	Human vs. ChatGPT Kappa	Human vs. ChatGPT P-value	ChatGPT vs. guidelines Kappa	ChatGPT vs. guidelines P-value
5-ASA	0.689	0.002	0.689	0.002	0.604	0.008
Steroids	0.406	0.028	0.578	0.005	0.771	0.001
Anti-TNF	0.289	0.073	0.345	0.047	0.685	0.003
Anti-integrins	0.486	0.013	0.336	0.05	0.771	0.001
Anti-IL-23	1	0	1	0	1	0
Thiopurines	0.441	0.054	0.313	0.161	0.771	0.001
Antibiotics	0.642	0.003	1	0	1	0
Diagnostic workup	1	0	1	0	1	0
Symptom management			1	0	1	0
Surgical consult			1	0	1	0
Continuous monitoring			1	0	1	0

5-ASA: 5-aminosalicylic acid; IL: interleukin; TNF: tumor necrosis factor.

ly in agreement but showed some minor variations.

Moderate to fair agreement ($\kappa \approx 0.3 - 0.5$)

Moderate to fair agreement was found in recommendations involving anti-TNF agents and anti-integrins. Kappa values ranged from 0.289 to 0.685 across evaluator pairs. There was noticeable variability in the recommendations for anti-TNF and anti-integrins, indicating differing clinical judgments or interpretations.

Low or non-significant agreement

The lowest agreement was observed for thiopurine therapy. While ChatGPT-5o and ECCO guidelines demonstrated substantial concordance ($\kappa = 0.771$), the comparison between human clinicians and ChatGPT-5o yielded a lower, non-significant agreement ($\kappa = 0.313$, $P = 0.161$).

Statistically significant agreement ($P < 0.05$) was observed in nearly all domains, with the exception of thiopurines in the human vs. ChatGPT-5o comparison (Table 3).

To complement the numerical analysis, Figure 1 displays a bar chart comparing agreement levels (Cohen's Kappa) across the 11 decision categories for each evaluator pair. Figure 2 provides a heatmap visualization of agreement intensity, with color-coded shading to highlight perfect, substantial, and variable concordance among ChatGPT-5o, human clinicians, and ECCO guidelines.

Discussion

This study evaluated the ability of ChatGPT-5o to generate clinical recommendations for IBD management that align with those of human providers and the ECCO guidelines [13, 14].

Findings were consistent with our earlier ChatGPT-4o analysis, but the updated 5o model provided greater stability in several therapeutic domains, underscoring the importance of ongoing evaluation as LLMs evolve.

Overall, ChatGPT-5o demonstrated a high level of concordance with both human providers' decisions and ECCO guidelines, especially in key areas such as antibiotic use, diagnostic workup, symptom management, surgical consult, and biologic therapy involving anti-IL-23 agents. This indicates its potential for standard IBD care [15]. These results suggest that ChatGPT-5o can reliably recommend foundational care components that are guideline-driven and less subject to clinical variation.

Areas of substantial agreement, such as the use of corticosteroids and 5-ASA, still reflect a strong alignment, suggesting that ChatGPT-5o generally follows the guidelines and human judgments closely. For example, Kappa values in the corticosteroid category were moderate to substantial, suggesting ChatGPT-5o appropriately identified when steroids were indicated but sometimes differed in recommendations, possibly due to the nuance required in patient-specific decisions.

Biologics like anti-TNF and anti-integrins demonstrated moderate to fair agreement, which suggests that ChatGPT-5o and humans may diverge in decision-making in more complex cases or where the guidelines leave room for interpretation, this may not fully reflect significant differences in therapeutic intent. In addition to guideline interpretive latitude, variability likely reflects the uptake of therapies not fully incorporated into ECCO at publication, including Janus kinase (JAK) inhibitors and S1P modulators, which can shift provider preferences in complex or refractory cases [16]. This highlights the need for AI to evolve alongside advancements in treatment modalities.

The lowest agreement was observed in thiopurine therapy, where human clinicians and ChatGPT-5o showed fair but statistically non-significant concordance. This suggests that ChatGPT-5o and human recommendations might differ sub-

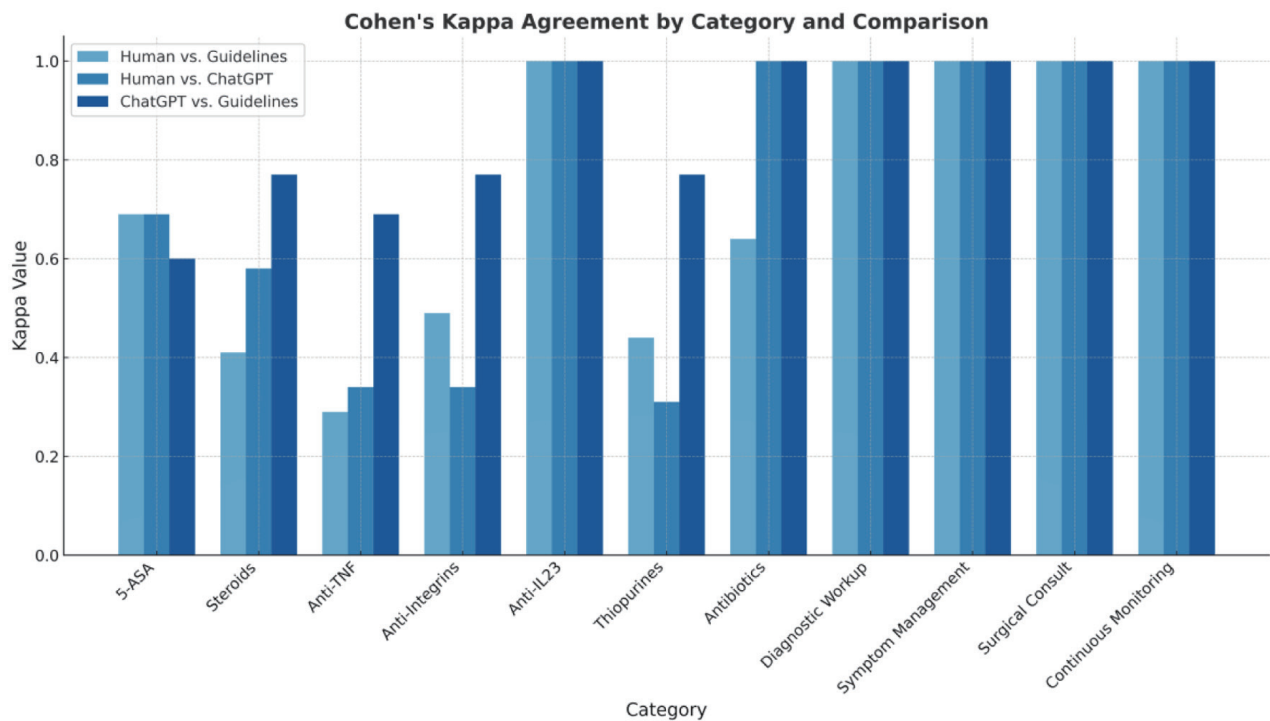


Figure 1. Bar chart for Cohen's Kappa values by category and comparison.

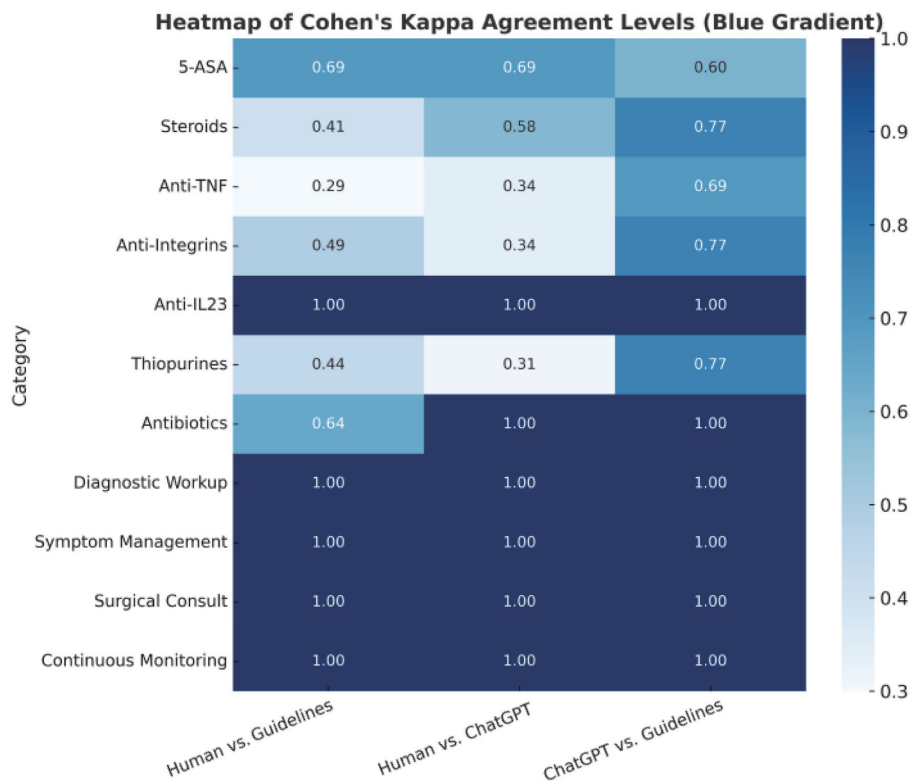


Figure 2. Cohen's Kappa agreement levels across different categories. The color scale represents the level of agreement: dark blue (1.0): perfect agreement (e.g., anti-IL-23, antibiotics, diagnostic workup, symptom management, surgical consultation, continuous monitoring). Medium blue ($\approx 0.6 - 0.8$): substantial agreement. Light blue ($\approx 0.3 - 0.5$): moderate to fair agreement, indicating some variability in alignment.

stantially in this area, potentially due to different thresholds for use or interpretations, decline in thiopurine use due to safety concerns or patient intolerance, and preference for early or first line treatment with biologics [16]. Importantly, ChatGPT-5o's recommendations were more closely aligned with ECCO guidelines in this area, highlighting the model's tendency to adhere to established protocols even when real-world clinical practice varies. These discrepancies may carry important clinical implications, as thiopurine initiation and monitoring are often tied to nuanced safety considerations, pharmacogenetics, and patient preferences. Therefore, over-reliance on guideline adherence without contextualization could risk suboptimal or unsafe recommendations in certain patients.

These findings support the potential of AI like ChatGPT-5o as supplementary tools in IBD management. ChatGPT-5o's strengths lie in its consistency, evidence alignment, and capacity to process structured clinical information to deliver guideline-adherent suggestions. However, areas of variability emphasize the importance of human oversight and clinical judgment, particularly in complex or evolving treatment landscapes [17]. Looking ahead, prospective studies should also evaluate how best to integrate AI-driven decision support into existing clinical workflows, ensuring that recommendations enhance, rather than disrupt, clinician-patient interactions.

Finally, while this study included a diverse and clinically relevant set of real-world cases, its generalizability is limited by the sample size and single-model assessment. This study was designed as a pilot exploratory analysis, and the 19 cases included were deliberately selected to represent diverse IBD presentations, enhancing its exploratory value. Because this study only evaluated a single AI model (ChatGPT-5o), it is unclear whether the findings would generalize to other LLMs. Future studies should include larger multi-center case series, additional disease scenarios (e.g., pediatric IBD, extraintestinal manifestations), and direct comparisons between multiple AI models (LLMs) to validate and expand these findings. Moreover, prospective integration into clinical workflows is necessary to evaluate the actual impact on decision-making, patient outcomes, and healthcare efficiency.

Ethical considerations of AI-supported IBD care

Even though our results show that ChatGPT-5o can follow guidelines and match clinical decisions in many cases, using AI in patient care brings up important ethical questions. These tools are not just technical; they affect how decisions are made, who makes them, and what role the patient plays.

One concern is responsibility. If an AI tool like ChatGPT makes a suggestion that causes harm, it is not clear who should be held accountable - the doctor, the company behind the model, or the hospital using it. Right now, there are no clear rules about this, and that creates uncertainty for clinicians and patients. Another issue is that ChatGPT may follow guidelines, but it does not understand the patient. It cannot see the bigger picture or adapt to individual needs the way a physician can. If doctors rely too much on the model, it could weaken shared decision-making, especially in cases where the

patient's values or unique circumstances matter more than the textbook answer. It cannot ask the patient how they are coping with their disease or what kind of support they have. These human experiences matter just as much as lab results, and they often determine whether treatment will succeed or fail. At the same time, AI could help improve access to IBD care in places where specialists are limited, which is an ethical opportunity that should not be ignored.

Lastly, even though we used only de-identified clinical information, it is still important to mention privacy and ethical oversight when AI is involved in research or care. Just because something is technically allowed does not mean we should ignore the ethical responsibility to protect patients and stay transparent.

Overall, AI tools like ChatGPT-5o have real potential to support IBD care, but they must be used carefully, with attention to their limits, and always under human supervision.

Conclusion

This study demonstrates that ChatGPT-5o can generate clinical recommendations for IBD that closely align with both human providers and the ECCO guidelines across multiple therapeutic and management domains. The model achieved perfect to substantial agreement in several areas governed by standardized protocols, including diagnostic evaluation, symptom management, surgical consultation, antibiotic use, biologic therapy with anti-IL-23, 5-ASA and corticosteroids, highlighting its potential to support evidence-based decision-making. Moderate to fair variability in categories such as anti-TNF and anti-integrins may reflect the inclusion of newer treatments, such as JAK inhibitors, which are not yet covered by ECCO guidelines. Thiopurine recommendations likely reflect the inherent complexity and evolving nature of IBD management, where clinical judgment and experience continue to play a central role.

These findings suggest that ChatGPT-5o may serve as a reliable adjunct in gastroenterology practice, particularly as a clinical decision-support tool for standardizing IBD care and improving guideline adherence, while addressing gaps in access to specialized care. However, its optimal use requires integration within systems that allow for expert oversight, contextual interpretation, and patient-centered adaptation.

At the same time, there are specific risks associated with over-reliance on AI, including the possibility of clinicians deferring too heavily to algorithmic outputs without considering patient nuance or emerging evidence. Responsible use therefore requires deliberate safeguards, including continuous clinician involvement, transparency about the model's limitations, and mechanisms for accountability when errors occur. Equally important is an interdisciplinary evaluation that brings together not only clinical expertise but also perspectives from bioethics and the health humanities. This broader lens helps ensure that patient dignity, autonomy, and lived experience remain central as AI is integrated into care.

In practice, clinicians might use AI most responsibly as a complement rather than a substitute: a way to cross-check decisions, enhance guideline adherence, and expand access, while ultimately maintaining the primacy of individualized

medical judgment. Further prospective studies are needed to validate these results across broader clinical scenarios, assess longitudinal outcomes, and determine how best to integrate such tools into routine clinical workflows without compromising safety or patient-centered care.

Supplementary Material

Suppl 1. Prompt and model details.

Acknowledgments

None to declare.

Financial Disclosure

This study received no external funding.

Conflict of Interest

The authors declare no conflict of interest.

Informed Consent

Not applicable. This study analyzed anonymized case summaries with no identifiable patient information.

Author Contributions

Daniel Aillaud-De-Uriarte: conceptualization, methodology, writing - original draft, project administration. Alejandro Cendejas-Higuera: data analysis, writing - review and editing. Luis A. Hernandez-Flores: supervision, critical revision, data analysis. Eric Acosta-Marquez: literature review, data analysis, review and editing. Monica Reyes-Bastidas and Haire Manzano-Cortes: independent evaluation of clinical cases, guideline validation, clinical insights - review and editing. All authors reviewed and approved the final manuscript and agree to be accountable for all aspects of the work.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declaration of Use of Generative AI and AI-Assisted Technologies

During the preparation of this work, the authors used ChatGPT-

5o (OpenAI, San Francisco, CA, USA) exclusively to generate clinical recommendations for evaluation within the study. The model was not used to write, edit, or format any portion of the manuscript. After using the tool, the authors reviewed and verified all generated material and take full responsibility for the content and integrity of the published article.

Abbreviations

AI: artificial intelligence; AGA: American Gastroenterological Association; 5-ASA: 5-aminosalicylic acid; CD: Crohn's disease; ECCO: European Crohn's and Colitis Organisation; EMR: electronic medical record; IBD: inflammatory bowel disease; IL: interleukin; IRB: Institutional Review Board; JAK: Janus kinase; LLM: large language model; NSAID: nonsteroidal anti-inflammatory drug; SPSS: Statistical Package for the Social Sciences; TNF: tumor necrosis factor; UC: ulcerative colitis

References

1. Kaplan GG. The global burden of IBD: from 2015 to 2025. *Nat Rev Gastroenterol Hepatol*. 2015;12(12):720-727. [doi pubmed](#)
2. Maloy KJ, Powrie F. Intestinal homeostasis and its breakdown in inflammatory bowel disease. *Nature*. 2011;474(7351):298-306. [doi pubmed](#)
3. Ng SC, Shi HY, Hamidi N, Underwood FE, Tang W, Benchimol EI, Panaccione R, et al. Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. *Lancet*. 2017;390(10114):2769-2778. [doi pubmed](#)
4. Mahid SS, Minor KS, Soto RE, Hornung CA, Galandiuk S. Smoking and inflammatory bowel disease: a meta-analysis. *Mayo Clin Proc*. 2006;81(11):1462-1471. [doi pubmed](#)
5. Rubin DT, Ananthakrishnan AN, Siegel CA, Sauer BG, Long MD. ACG clinical guideline: ulcerative colitis in adults. *Am J Gastroenterol*. 2019;114(3):384-413. [doi pubmed](#)
6. Romero RK, Magro DO, Queiroz NSF, Damiao A, Teixeira FV, Nones RB, Sassaki LY, et al. Perception and clinical decisions from inflammatory bowel diseases' specialists towards positioning of new therapies in Crohn's disease and ulcerative colitis: A national web-based survey from the Brazilian IBD study group (GEDIIB). *Gastroenterol Hepatol*. 2022;45(7):499-506. [doi pubmed](#)
7. Singh A, Mahajan R, Kedia S, Dutta AK, Anand A, Bernstein CN, Desai D, et al. Use of thiopurines in inflammatory bowel disease: an update. *Intest Res*. 2022;20(1):11-30. [doi pubmed](#)
8. Kuo CY, Chiu HM. Application of artificial intelligence in gastroenterology: Potential role in clinical practice. *J Gastroenterol Hepatol*. 2021;36(2):267-272. [doi pubmed](#)
9. Zand A, Stokes Z, Sharma A, van Deen WK, Hommes

- D. Artificial intelligence for inflammatory bowel diseases (IBD); Accurately predicting adverse outcomes using machine learning. *Dig Dis Sci.* 2022;67(10):4874-4885. [doi pubmed](#)
10. Morley J, Machado CCV, Burr C, Cowls J, Joshi I, Taddeo M, Floridi L. The ethics of AI in health care: A mapping review. *Soc Sci Med.* 2020;260:113172. [doi pubmed](#)
 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 pubmed](#)
 12. Cross JL, Choma MA, Onofrey JA. Bias in medical AI: Implications for clinical decision-making. *PLOS Digit Health.* 2024;3(11):e0000651. [doi pubmed](#)
 13. Raine T, Bonovas S, Burisch J, Kucharzik T, Adamina M, Annese V, Bachmann O, et al. ECCO guidelines on therapeutics in ulcerative colitis: medical treatment. *J Crohns Colitis.* 2022;16(1):2-17. [doi pubmed](#)
 14. Torres J, Bonovas S, Doherty G, Kucharzik T, Gisbert JP, Raine T, Adamina M, et al. ECCO guidelines on therapeutics in Crohn's disease: medical treatment. *J Crohns Colitis.* 2020;14(1):4-22. [doi pubmed](#)
 15. Da Rio L, Spadaccini M, Parigi TL, Gabbiadini R, Dal Buono A, Busacca A, Maselli R, et al. Artificial intelligence and inflammatory bowel disease: Where are we going? *World J Gastroenterol.* 2023;29(3):508-520. [doi pubmed](#)
 16. D'Amico F, Peyrin-Biroulet L, Danese S. JAK or GUT selectivity: tipping the balance for efficacy and safety in ulcerative colitis. *J Crohns Colitis.* 2020;14(9):1185-1187. [doi pubmed](#)
 17. Fast D, Adams LC, Busch F, Fallon C, Huppertz M, Siepmann R, Prucker P, et al. Autonomous medical evaluation for guideline adherence of large language models. *NPJ Digit Med.* 2024;7(1):358. [doi pubmed](#)