United European Gastroenterology Journal

WILEY

UNITED EUROPEAN GASTROENTEROLOGY
ueg journal

ORIGINAL ARTICLE OPEN ACCESS

Inter-Rater Disagreements in Applying the Montreal Classification for Crohn's Disease: The Five-Nations Survey Study

Offir Ukashi[1,2] | Aurelien Amiot[3] | David Laharie[4] | Luis Menchén[5,6] | Ana Gutiérrez[7] | Samuel Fernandes[8,9,10] | Tommaso Pessarelli[11] | Fábio Correia[12] | Carlos Gonzalez-Muñoza[13,14] | Julia López-Cardona[15] | Giulio Calabrese[16] | Rocio Ferreiro-Iglesias[17,18] | Natalie Tamir-Degabli[2,19] | Nikolas Konstantine Dussias[20,21] | Amjad Mousa[22] | Raquel Oliveira[23] | Nicolas Richard[24] | Ido Veisman[1,2] | Kassem Sharif[1,2] | Shomron Ben-Horin[1,2] | Carlos Soutullo-Castiñeiras[25,26] | Gabriele Dragoni[27] | Silvia Rotulo[28] | Agnese Favale[29] | Louis Calméjane[30] | Thomas Bazin[31,32] | Alfonso Elosua[33] | Sara Lopes[34] | Carla Felice[35] | Violeta Mauriz[17] | Inês Coelho Rodrigues[8,9] | Julia Jougon[36] | Inês Botto[8,9] | Helena Tavares de Sousa[23] | Lorenzo Bertani[37] | Paula Ripoll Abadía[38] | Alice De Bernardi[39] | Yamile Zabana[40] | Xavier Serra-Ruiz[41] | Anna Viola[42] | Manuel Barreiro-de Acosta[17,18] | Henit Yanai[2,19] | Alessandro Armuzzi[43,44] | Fernando Magro[45] | Uri Kopylov[1,2]

[1]Sheba Medical Center, Institute of Gastroenterology, Ramat-Gan, Israel | [2]Faculty of Medical and Health Sciences, Tel-Aviv University, Tel-Aviv, Israel | [3]Department of Gastroenterology, Hopitaux Universitaires Bicêtre, AP-HP, Université Paris Saclay, INSERM U1018 CESP, Le Kremlin Bicêtre, France | [4]CHU de Bordeaux, Centre Medico-Chirurgical Magellan, Gastroenterology Department, Hôpital Haut-Lévêque, Université de Bordeaux, INSERM CIC 1401, Bordeaux, France | [5]Hospital General Universitario - Instituto de Investigación Sanitaria Gregorio Marañón, Madrid, Spain | [6]Departamento de Medicina, Universidad Complutense, Madrid, Spain | [7]Gastroenterology Department Hospital General Universitario Dr Balmis of Alicante, ISABIAL, Centro de Investigación Biomédica en Red en Enfermedades Hepáticas y Digestivas (CIBERehd), Alicante, Spain | [8]Gastroenterology and Hepatology Department, Unidade Local de Saúde Santa Maria, Lisbon, Portugal | [9]Clínica Universitária de Gastrenterologia, Faculdade de Medicina de Lisboa, Lisboa, Portugal | [10]Grupo de estudos de Doenças Inflamatórias do Intestino (GEDII), Porto, Portugal | [11]Department of Pathophysiology and Transplantation, University of Milan, Milan, Italy | [12]Gastroenterology Department, Prof. Dr. Fernando Fonseca Hospital, Amadora, Portugal | [13]H. Santa Creu i Sant Pau (Gastroenterology Department), Barcelona, Spain | [14]Departament de Medicina, Universitat Autònoma de Barcelona, Barcelona, Spain | [15]Gastroenterology and Hepatology Department, University Hospital Ramon y Cajal, Madrid, Spain | [16]Gastroenterology Unit, Clinical Medicine and Surgery Department, University of Naples Federico II, Naples, Italy | [17]Gastroenterology Department Hospital Universitario de Santiago de Compostela, A Coruña, Spain | [18]Fundacion Instituto de Investigación Sanitaria de Santiago de Compostela (FIDIS), A Coruña, Spain | [19]Division of Gastroenterology, Rabin Medical Center, Petah Tikva, Israel | [20]IBD Unit, IRCCS Azienda Ospedaliero-Universitaria di Bologna, Bologna, Italy | [21]Department of Medical and Surgical and Sciences, University of Bologna, Bologna, Italy | [22]Gastroenterology Department, Bnai Zion Medical Center, Haifa, Israel | [23]Gastroenterology Department, Unidade Local de Saúde do Algarve, Portimão, Portugal | [24]"Nutrition, Inflammation, and Microbiota-Gut-Brain Axis," CHU Rouen, Department of Gastroenterology, University of Rouen Normandie, INSERM, Normandie University, ADEN UMR1073, Rouen, France | [25]Gastrointestinal Endoscopy Research Group, Health Research Institute Hospital La Fe (IISLaFe), Valencia, Spain | [26]Gastrointestinal Endoscopy Unit, Hospital Universitari I Politècnic La Fe, Valencia, Spain | [27]IBD Referral Centre, Clinical Gastroenterology Unit, Careggi University Hospital, Florence, Italy | [28]Department of Maternal and Child Health, Pediatric Gastroenterology and Liver Unit, Umberto I Hospital, Sapienza University of Rome, Rome, Italy | [29]Department of Medical Science and Public Health, University of Cagliari, Cagliari, Italy | [30]Université Paris Cité, Paris, France | [31]Department of Gastroenterology and Nutritional Support, Center for Intestinal Failure, Reference Centre of Rare Disease MarDI, Assistance Publique—Hôpitaux de Paris (AP-HP) Beaujon Hospital, University Paris Cité, Clichy, France | [32]Infection & Inflammation, Unité Mixte de Recherche (UMR) 1173, Inserm, Université de Versailles—Saint-Quentin-en-Yvelines (UVSQ)/Université Paris Saclay, Montigny-le-Bretonneux, France | [33]Gastroenterology Department, Hospital Universitario de Navarra, IdiSNA, Navarra Institute for Health Research, Pamplona, Spain | [34]Unidade Local de Saúde da Arrábida, Setúbal, Portugal | [35]Department of Medicine, University of Padova, Padova, Italy | [36]Hepatogastroenterology Department, University of Lille, Lille, France | [37]Tuscany North West ASL, Department of Internal Medicine, Pontedera Hospital, Pontedera, Italy | [38]Gastroenterology, Hospital Universitario y Politécnico La Fe, Valencia, Spain | [39]IBD Center, Gastroenterology Unit, Rho Hospital, ASST Rhodense, Rho, Italy | [40]Hospital Universitari Mútua Terrassa and Centro de Investigación Biomédica en Red en Enfermedades Hepáticas y Digestivas (CIBERehd), Terrassa, Spain | [41]Crohn's and Colitis Unit, Gastroenterology Department, Hospital Universitari Vall d'Hebron, Barcelona, Spain | [42]IBD-Unit, Department of Clinical and Experimental Medicine, University of Messina, Messina, Italy | [43]IBD Center, IRCCS Humanitas Research Hospital, Milan, Italy | [44]Department of Biomedical Sciences, Humanitas University, Milan, Italy | [45]CINTESIS@RISE, Faculty of Medicine, The University of Porto, Porto, Portugal

Correspondence: Offir Ukashi (offirukashi@gmail.com)

Received: 24 September 2024 | **Revised:** 1 November 2024 | **Accepted:** 5 November 2024

Funding: The authors received no specific funding for this work.

Keywords: complicated disease phenotype | Crohn's disease | inflammatory bowel diseases | large language models | montreal classification

ABSTRACT

Background: The Montreal classification has been widely used in Crohn's disease since 2005 to categorize patients by the age of onset (A), disease location (L), behavior (B), and upper gastrointestinal tract and perianal involvement. With evolving management paradigms in Crohn's disease, we aimed to assess the performance of gastroenterologists in applying the Montreal classification.

Methods: An online survey was conducted among participants at an international educational conference on inflammatory bowel diseases. Participants classified 20 theoretical Crohn's disease cases using the Montreal classification. Agreement rates with the inflammatory bowel diseases board (three expert gastroenterologists whose consensus rating was considered the gold standard) were calculated for gastroenterologist specialists and fellows/specialists with $\leq$ 2 years of clinical experience. A majority vote < 75% among participants was considered a notable disagreement. The same cases were classified using three large language models (LLMs), ChatGPT-4, Claude-3, and Gemini-1.5, and assessed for agreement with the board and gastroenterologists. Fleiss Kappa was used to assess within-group agreement.

Results: Thirty-eight participants from five countries completed the survey. In defining the Montreal classification as a whole, specialists (21/38 [55%]) had a higher agreement rate with the board compared to fellows/young specialists (17/38 [45%]) (58% vs. 49%, $p = 0.012$) and to LLMs (58% vs. 18%, $p < 0.001$). Disease behavior classification was the most challenging, with 76% agreement among specialists and fellows/young specialists and 48% among LLMs compared to the inflammatory bowel diseases board. Regarding disease behavior, within-group agreement was moderate (specialists: $k = 0.522$, fellows/young specialists: $k = 0.532$, LLMs: $k = 0.577$; $p < 0.001$ for all). Notable points of disagreement included: defining disease behavior concerning obstructive symptoms, assessing disease extent via video capsule endoscopy, and evaluating treatment-related reversibility of the disease phenotype.

Conclusions: There is significant inter-rater disagreement in applying the Montreal classification, particularly for disease behavior in Crohn's disease. Improved education or revisions to phenotype criteria may be needed to enhance consensus on the Montreal classification.

1 | Introduction

Crohn's disease (CD) is a chronic inflammatory condition that can affect the entire gastrointestinal tract, leading to tissue damage and complications such as fistulas, abscesses, and bowel obstructions [1]. At diagnosis, 26% of CD patients present with a complicated disease phenotype, increasing to 48% at 5 years and 70% at 10 years [2]. Early age at diagnosis [3–5], fibro-stenotic [3, 5] or penetrating [5, 6] disease behavior and involvement of the proximal small bowel (SB) [7] have been proven to predict CD-related complications and future intestinal resections.

The Montreal classification [8, 9], developed in 2005 as a revision of the Vienna classification [10], is the most widely used system for CD. It includes three key parameters: age at onset, disease extent, and behavior, with two modifiers for upper gastrointestinal tract (GIT) and perianal involvement. Disease behavior is considered permanent unless progression occurs, and it is recommended to determine it no earlier than 5 years post-diagnosis [9]. A recently published study [11] showed that using a revised bidirectional version of the Montreal classification reduced the number of patients classified with complicated disease (B2/3) at 5 years to 10%, compared to 42% with the traditional unidirectional system. Notably, patients with complicated CD have a higher risk of disease-related complications and increased lifetime medication use [3, 6, 12]. Thus, a more intensive management approach may be warranted, making uniform and consistent disease classification among gastroenterologists essential [12].

Since the introduction of the Montreal classification [9], treatment and monitoring paradigms for CD have evolved significantly. Previously, clinical features and ileo-colonoscopy findings were the primary basis for classification as imaging studies were less accessible. However, ileo-colonoscopy may miss up to 20% of cases due to technical issues that preclude ileal intubation [13] or diseases beyond its reach [14]. Today, computed tomography enterography (CTE), magnetic resonance enterography (MRE), intestinal ultrasound (IUS), and video capsule endoscopy (VCE) are widely used and may influence disease classification. Additionally, advanced therapies like biologics have shown potential in reversing bowel damage [15, 16].

Large Language Models (LLMs) are artificial intelligence systems trained on extensive text data to generate natural language responses, aiding in patient education and supporting rare condition diagnoses [17]. However, their clinical application remains limited, particularly in complex cases requiring human

Summary

- Summary of the established knowledge on this subject
  - The Montreal classification, introduced in 2005 as a revision of the Vienna classification (1998), is the most widely used classification system for CD.
  - Since the Vienna and Montreal classifications, treatment and monitoring paradigms for CD have evolved significantly.
  - Initially, classification relied on clinical features and ileo-colonoscopy findings as imaging studies were less accessible.
  - Today, advanced imaging techniques like magnetic resonance enterography, intestinal ultrasound, and video capsule endoscopy (VCE) are widely used and may influence disease classification.
  - Large language models (LLMs) show potential for patient education and aiding in the diagnosis of rare conditions by generating quick, natural language responses.
- Significant/new Findings of This Study
  - Specialists in gastroenterology (> 2 years of experience) showed higher agreement with the IBD board in accurately applying the Montreal classification than fellows/young specialists (58% vs. 49%, $p = 0.012$) and LLMs (58% vs. 18%, $p < 0.001$).
  - Disease behavior classification was the most challenging, with 76% agreement among specialists/fellows and 48% among LLMs, compared to the IBD board.
  - Major points of disagreement involved defining disease location using VCE and assigning B2 disease behavior based on clinical, radiologic, and endoscopic features.

expertise, and their effectiveness depends on the quality of their training data [17].

In this study, we aimed to examine the current practical relevance of the Montreal classification among gastroenterologists with varying levels of expertise in various countries, as well as its use by LLMs.

2 | Materials and Methods

2.1 | Study Design and Participants

Twenty theoretical CD vignettes were created (Supporting Information S1), and a Google Form survey was developed. The survey collected data on country, training level (fellow/specialist with ≤ 2 years of experience or specialist with > 2 years of experience), gastroenterology interest (inflammatory bowel disease [IBD] versus non-IBD), practice type (academic, private, community), healthcare setting (primary, secondary, tertiary), practice location (urban/rural), involvement in clinical trials, and resource availability. Participants were asked to assign the Montreal classification for each case, including age at onset (A1, A2, A3), location (L1, L2, L3), disease behavior (B1, B2, B3), and modifiers for upper GIT (L4) and perianal involvement (P).

On July 10th, 2024, 128 gastroenterologists who participated in the Five Nations Conference—an international educational IBD conference held in Porto, Portugal, from May 30th to June 2nd, 2024—received an email invitation to participate in this survey study. The email included a link to the survey. The five participating nations were Italy, France, Portugal, Spain, and Israel. The survey was available until July 20th, 2024. In case participants missed any questions, or any component of the Montreal classification, or made double selections in any clinical vignette, we sent an email after July 20th with the relevant questions asking them to complete or amend the answers accordingly (only technical amendments as described above were allowed). For transparency and clarity, the CHERRIES guidelines checklist is provided in the Supporting Information S1 and includes data regarding the study survey design, development, participant incentives, and security and protection measures.

2.2 | IBD Board

Three IBD experts formed the IBD board (FM, AA, HY). In cases of discrepancies in the Montreal classification across any of the clinical vignettes, the board reached a consensus through a majority vote following a consensus meeting discussion. This consensus was established as the gold standard for the study.

2.3 | Gastroenterologist Groups

The participating gastroenterologists were divided into two groups based on their experience. The “fellows/young specialists' group” included gastroenterologists during their fellowship or training and specialists with up to 2 years of experience. The “specialists' group” included gastroenterologists with more than 2 years of experience.

2.4 | Large Language Models

The LLMs group included ChatGPT-4 (OpenAI. [2024], https://chatgpt.com), Claude-3 (Anthropic. [2023]. Claude [Artificial intelligence]. https://claude.ai/), and Gemini-1.5 (Google [2023], https://gemini.google.com), which were accessed between July 10 and 29, 2024. To assess the LLMs' ability to apply the Montreal classification system, we used a two-part prompt sequence. The initial prompt “Are you aware of the Montreal classification?” was followed by “Can you provide the Montreal classification for these 20 cases?” along with 20 clinical cases from the Supporting Information S1. This standardized prompt sequence was applied consistently across all tested LLMs. ChatGPT-4 and Claude-3 provided classifications for all clinical vignettes at once, except in a single instance (Case 3) where Claude-3 did not provide the B classification. This case was then submitted separately, and an eligible answer was obtained. For Gemini-1.5, due to input limitations, we had to modify our approach by presenting the clinical vignettes individually rather than as a complete set. Each case was submitted separately to obtain its Montreal classification, in contrast to other LLMs which could process all 20 cases simultaneously. Despite this, the Montreal classification could not be assigned by Gemini-1.5 in seven out of 20 cases, with the

classification being deemed not applicable for these specific cases. Only one submission was made using the LLMs, except for a single case with Claude-3, as mentioned above. The complete outputs from all tested LLMs in response to the input prompts are provided in the Supporting Information S1. A representative example of the interaction with ChatGPT-4 is illustrated in Figure S1, which shows a snapshot of the chat interface and responses.

2.5 | Study Objectives

The following objectives were established for this survey study:

- To examine the agreement rate between the IBD board and each gastroenterologist group in assigning the Montreal classification.
- To compare the performance of the gastroenterologist groups in assigning the Montreal classification using the IBD board as the gold standard.
- To examine the combined agreement rate of the LLMs with the IBD board and to compare its performance with the gastroenterologist groups.
- To assess the agreement within each study group in defining each of the components of the Montreal classification.
- To highlight the notable points of disagreement in classifying CD patients based on the Montreal classification, defining it as less than 75% agreement among the survey participants.

2.6 | Statistical Analysis

Dichotomous variables were expressed as proportions. The agreement rate between the study groups and the IBD board is presented as proportions. Comparisons between the study groups were performed using Fisher's exact test, examining the Montreal classification as a whole and for each of its components. Agreement within groups was evaluated using the Fleiss kappa test for Montreal's components. The level of agreement was classified as follows: less than 0—poor; 0.01–0.20—slight; 0.21–0.40—fair; 0.41–0.60—moderate; 0.61–0.80—substantial; 0.81–1.00—almost perfect [18]. Fleiss' kappa agreement was reported along with 95% confidence intervals (CI) to indicate the effect sizes of the results. We had two out of 760 (0.2%) clinical cases that were not completed and 16 out of 3800 (0.4%) Montreal components with double selections. These data were excluded before data analysis. All statistical tests were two-sided, and a p-value of < 0.05 was considered statistically significant. Statistical analyses were performed using SPSS software (IBM SPSS Statistics for Windows, Version 26; IBM Corp., Armonk, NY, USA, 2019).

2.7 | Study Ethics

Since all clinical vignettes in this survey were theoretical, neither Helsinki approval nor patient consent was required.

3 | Results

3.1 | Survey Participants

Out of 128 invited gastroenterologists, 41 (32%) completed the survey. Three were considered as IBD board members. Among the remaining participants ($n = 38$), 21 (55%) were specialists with over 2 years of experience, while 17 (45%) were in training or had up to 2 years of experience. Figure 1 demonstrates the characteristics of the study participants.

3.2 | Clinical Vignettes

Table 1 details the patients' baseline characteristics, the Montreal classification (as determined by the IBD board), and the distribution of diagnostic tools used in the clinical vignettes. Out of 20 cases, 17 (85%) cases used ileo-colonoscopy, followed by 14 (70%) cases used abdominal cross-sectional imaging, and 4 (20%) cases used SB-VCE.

The distribution of the Montreal classification, as assigned by the IBD board, is shown in Table 1. Eleven (55%) cases were classified as having inflammatory (B1) disease behavior, followed by six (30%) cases of stricturing (B2) disease behavior and three (15%) cases of penetrating (B3) disease behavior. Four (20%) cases were classified as having perianal disease involvement, and six (30%) cases were classified as having upper GIT involvement.

3.3 | Agreement Rates in the Gastroenterologist Groups

Using the IBD board as a reference, the specialists had 58% of correct Montreal classifications compared to 49% for the fellows/young specialists ($p = 0.012$). As depicted in Figure 2, the specialists had a significantly better performance in classifying disease anatomic location (89% vs. 79%, $p < 0.001$) than the latter. For all other components of the Montreal classification, their performance was comparable. Notably, correctly classifying disease behavior was the most challenging, with only a 76% agreement rate observed in both gastroenterologist groups.

3.4 | Gastroenterologist Within-Group Agreement

As detailed in Table 2, the agreement among the specialists was almost perfect in classifying the age at onset ($k = 0.823$, 95% CI 0.822–0.823, $p < 0.001$) and perianal disease involvement ($k = 0.902$, 95% CI 0.901–0.903, $p < 0.001$) and substantial for defining disease anatomic location ($k = 0.749$, 95% CI 0.748–0.750, $p < 0.001$) and upper GIT involvement ($k = 0.655$, 95% CI 0.654–0.656, $p < 0.001$). There was only a moderate agreement among specialists in classifying disease behavior ($k = 0.522$, 95% CI 0.521–0.522, $p < 0.001$). There was moderate agreement among fellow/young specialists in defining disease anatomic extent ($k = 0.573$, 95% CI 0.572–0.573, $p < 0.001$), disease behavior ($k = 0.532$, 95% CI 0.531–0.533 $p < 0.001$), and upper GIT involvement ($k = 0.550$, 95% CI 0.549–0.551, $p < 0.001$).

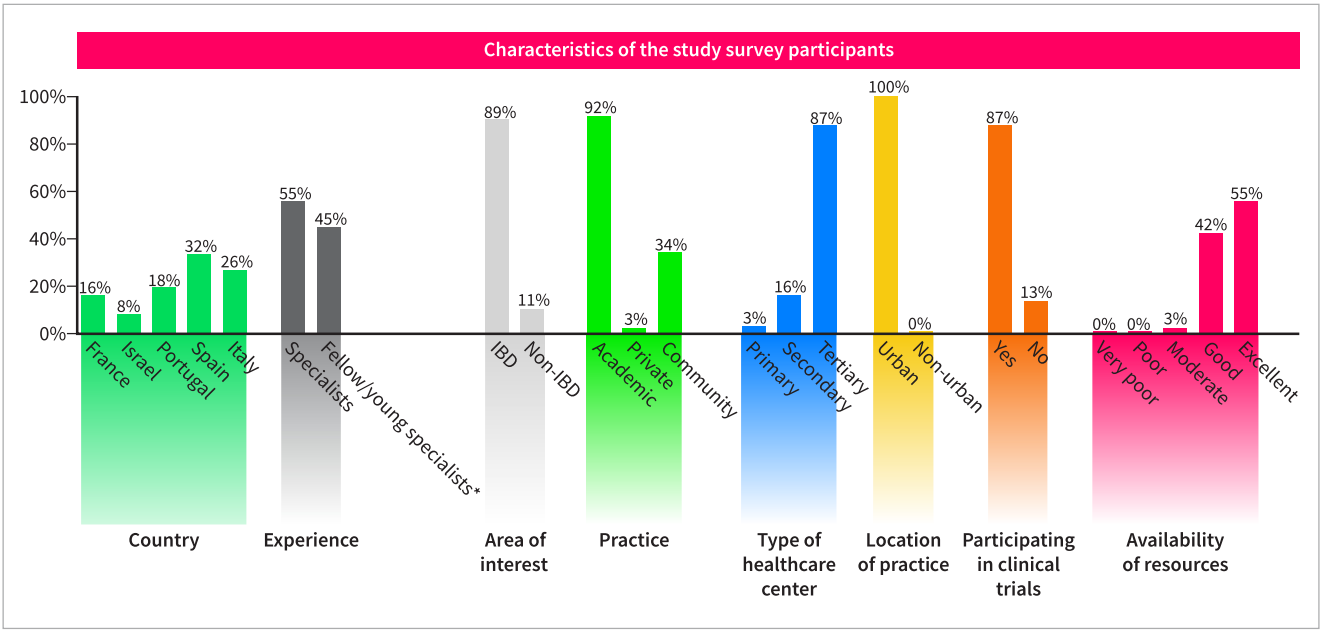

FIGURE 1 | Characteristics of the study survey participants. *Specialists with up to 2 years of experience. #Multiple choices were allowed for assigning the type of practice and healthcare center.

3.5 | LLMs' Agreement Rate

The LLMs had an 18% agreement rate in correctly classifying the Montreal classification of the study clinical vignettes. The performance of LLMs in defining the Montreal classification was significantly inferior to that of both gastroenterologist groups, except for a numerically higher performance rate of the fellows/young specialists regarding disease anatomic extent (Figure 2).

As depicted in Figure 3, the performance of ChatGPT-4 and Claude-3 in defining the Montreal classification was comparable. Gemini-1.5 had the poorest performance among the LLMs examined in this study, with significantly lower accuracy in defining disease location (50% vs. 85%, $p = 0.04$), upper GIT involvement (40% vs. 80%, $p = 0.02$), and perianal involvement (50% vs. 100%, $p < 0.001$) compared to ChatGPT-4, and significantly poorer performance compared to Claude-3 in classifying perianal involvement (50% vs. 95%, $p = 0.003$).

In a sensitivity analysis excluding Gemini-1.5, LLMs still performed significantly worse than both gastroenterologist groups in defining the Montreal classification, especially in classifying age at onset and disease behavior (compared to specialists only). However, they showed comparable performance in classifying upper GIT and perianal involvement (Supporting Information S1: Table S1).

3.6 | LLMs' Within-Group Agreement

While the agreement within LLMs was moderate regarding age at onset, disease anatomic location, and disease behavior, there was a poor agreement regarding upper GIT involvement ($k = -0.036$, 95% CI −0.043 to −0.030, $p = 0.726$) and slight agreement regarding perianal disease ($k = 0.179$, 95% CI 0.172 to 0.185, $p = 0.068$), as presented in Table 2.

3.7 | Notable Points of Disagreement in Defining Montreal classification

There were nine cases with less than 75% agreement on classifying disease behavior (B) and four cases with less than 75% agreement on classifying disease location (L) and upper GIT involvement (L4). Details on notable points of disagreement and the IBD board classification are presented in Table 3. Key disagreements included: assigning disease extent in the proximal and middle SB tertiles using VCE, defining B2 disease behavior when the ileocecal valve (ICV) could not be intubated or when VCE/Patency capsule (PC) retention occurred without strictures on cross-sectional imaging, determining if fistulizing perianal disease is B3 disease behavior, and whether jejunal disease qualifies as upper GIT involvement.

4 | Discussion

In this survey, we assessed the performance of the Montreal classification of CD among gastroenterologists from various countries and expertise levels. To our knowledge, this is the first study to examine its practical relevance in the era of advanced therapies and diagnostics such as MRE, VCE, and IUS. Disease behavior (B) was the most challenging to define, showing the lowest agreement with the IBD board and only moderate agreement within the survey participants, highlighting the need to revise the classification to align with current management paradigms. This is also the first study to assess LLMs in applying the Montreal classification, revealing their poorer performance

TABLE 1 | Patients' baseline characteristics, the Montreal classification (as determined by the IBD board), and the distribution of diagnostic tools used in the clinical vignettes.

	***N* = 20 (100%)**
Demographics	
Age, median (IQR)	29.5 (22.0–42.5)
Male, *n* (%)	7 (35%)
Montreal classification	
Age at onset (A), *n* (%)	
A1 (≤ 16 years-old)	2 (10%)
A2 (17–40 years-old)	14 (70%)
A3 (> 40 years-old)	4 (20%)
Location (L), *n* (%)	
L1(Ileal)	9 (45%)
L2 (Colonic)	5 (25%)
L3 (Ileo-colonic)	4 (20%)
Disease behavior (B), *n* (%)	
B1(non-stricturing, non-penterating)	11 (55%)
B2 (stricturing)	6 (30%)
B3 (penetrating)	3 (15%)
Upper GIT involvement (L4), *n* (%)	6 (30%)
Perianal involvement (P), *n* (%)	4 (20%)
Inflammatory biomarkers [a]	
CRP mg/dL, median (IQR)	4 (0.35–6.75)
FC μg/g, median (IQR)	825 (248–1025)
Usage distribution of diagnostic tools	
Ileo-colonoscopy	17 (85%)
CTE	5 (25%)
MRE	9 (45%)
Gastroscopy	3 (15%)
Pelvic MRI	2 (10%)
IUS	1 (5%)
VCE	4 (20%)

Abbreviations: CRP, C-reactive protein; CTE, computed tomography enterography; FC, fecal calprotectin; GIT, gastrointestinal tract; IQR, interquartile range; IUS, intestinal ultrasound; MRE, magnetic resonance enterography; MRI, magnetic resonance imaging; VCE, video capsule endoscopy.
[a]In four cases, inflammatory biomarkers were not provided.

compared to gastroenterologists, underscoring the complexity of accurate application.

Patients with CD who have a complicated disease phenotype are at a higher risk of experiencing disease-related complications, including longer hospital stays, abdominal surgeries, and increased medication use throughout their lives [3, 6, 12]. Consequently, it is reasonable to consider that physicians might opt for a more intensive management approach for these patients [12]. In this survey study, we observed considerable disagreement rates across all study groups compared with the IBD board when assigning disease behavior. Additionally, the within-group agreement on disease behavior assignment was only moderate across all study groups, aligning with findings from previously published studies evaluating the performance of the Montreal classification among gastroenterologists [19–21].

In this study, we created theoretical clinical vignettes to reflect the wide range of patients with CD. However, we likely did not encompass the entire spectrum of disease presentation in this population. According to the Vienna classification, stricturing disease is defined as persistent luminal narrowing, as determined by radiologic, endoscopic, or surgical findings, combined with pre-stenotic dilation and/or signs or symptoms of obstruction [10]. The IBD board adhered to the original classification criteria that emphasized the clinical presentation over imaging or endoscopy, assigning stricturing disease behavior to patients following VCE/PC retention only if they exhibited prior obstructive symptoms. Additionally, patients with endoscopic ICV stenosis that precluded intubation but had non-stenotic findings on radiologic imaging were classified as having inflammatory disease behavior. These cases led to notable disagreement among the survey participants in assigning B1 versus B2 disease behavior. Notably, according to the CONSTRICT consensus and the STAR Consortium, inability to pass an adult colonoscope through a narrowed area is sufficient to define a luminal stricture, and obstructive symptoms are not required for stricture definition [22, 23].

VCE retention can cause SB obstruction and may require surgery for symptomatic strictures, but its definition is based on an arbitrary 2-week retention period, regardless of obstructive symptoms [24]. Similarly, PC retention, or "unpassed PC," is defined by 30–33 hour timeframe or detection on abdominal X-ray, which has only a 50% accuracy rate in locating the capsule within the SB [24]. This timeframe can be affected by delayed colonic transit or constipation [25]. However, PC retention has been shown to predict poor outcomes in quiescent CD, including increased hospitalizations and the need for surgery [26]. VCE and PC, introduced after the Vienna and Montreal classifications, are important for diagnosing and monitoring CD but may have varying levels of familiarity and availability worldwide, affecting experience. The increased use of MRE also underscores the need to reassess and refine disease classification criteria to align with current diagnostics. However, these tools present challenges in assigning disease behavior and location, as our findings show. Accurate definitions are essential to avoid misclassification.

The appropriate timeframe for accurately classifying the CD phenotype remains controversial [9]. While 26% of CD patients are diagnosed with B2/B3 disease at diagnosis, this rate rises significantly over time, reaching 48% by 5 years and 70% by 10 years post-diagnosis [2]. A 5-year timeframe is recommended for more accurate classification, but this can leave many CD patients unclassified early on [9]. Although disease behavior is not the sole factor, initial classification can significantly impact the decision to implement intensive treatment strategies, which are crucial for prognosis [27]. A recent study by Bokemeyer et al. found that a revised bidirectional Montreal classification identified only 10% of patients with a complicated disease phenotype

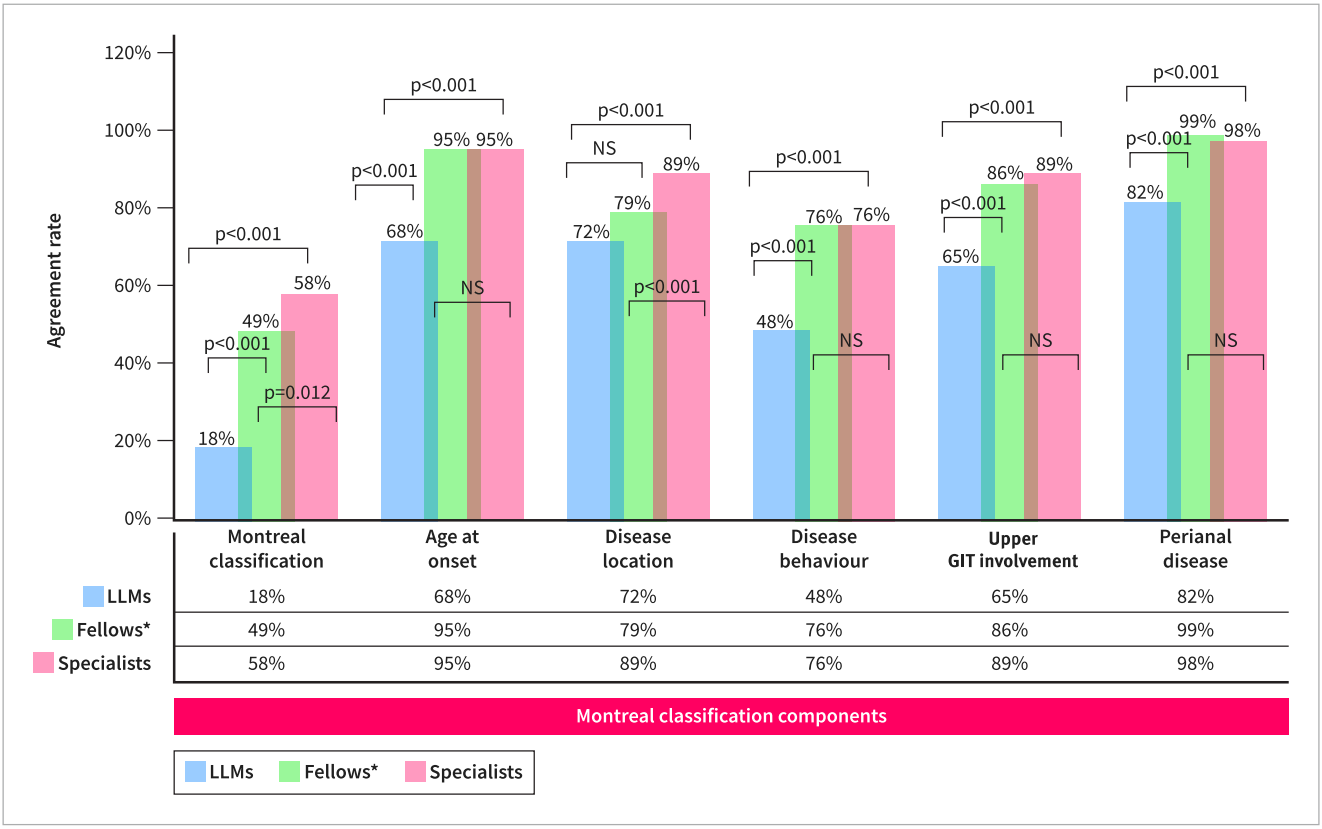

FIGURE 2 | The agreement rates with the inflammatory bowel disease board's consensus among the study groups. GIT, gastrointestinal tract; LLMs, language learning models. *Including young specialists with up to 2 years of experience.

TABLE 2 | The within-group agreement regarding the components of the Montreal classification.

	LLM's			Fellows/young specialists			Specialists		
	Fleiss kappa	95% CI	*p*-value	Fleiss kappa	95% CI	*p*-value	Fleiss kappa	95% CI	*p*-value
Age at onset (A)	0.513	0.508 to 0.518	< 0.001	0.817	0.816 to 0.818	< 0.001	0.823	0.822 to 0.823	< 0.001
Location (L)	0.528	0.524 to 0.533	< 0.001	0.573	0.572 to 0.573	< 0.001	0.749	0.748 to 0.750	< 0.001
Disease behavior (B)	0.577	0.572 to 0.582	< 0.001	0.532	0.531 to 0.533	< 0.001	0.522	0.521 to 0.522	< 0.001
Upper GIT involvement (L4)	−0.036	−0.043 to −0.030	0.726	0.550	0.549 to 0.551	< 0.001	0.655	0.654 to 0.656	< 0.001
Perianal disease (P)	0.179	0.172 to 0.185	0.068	0.930	0.929 to 0.931	< 0.001	0.902	0.901 to 0.903	< 0.001

Note: The level of agreement was classified as follows: less than 0—poor; 0.01–0.20—slight; 0.21–0.40—fair; 0.41–0.60—moderate; 0.61–0.80—substantial; 0.81–1.00—almost perfect [17].
Abbreviations: CI, confidence interval; LLMs, Large language models.

at 5 years, compared to 42% with the traditional unidirectional classification [11]. In this context, the IBD board strongly supported classifying patients with previously documented stricturing or penetrating complications as having a B2/B3 disease behavior, regardless of their current clinical status. This issue led to notable disagreement among the survey participants. CREOLE and STRIDENT studies have demonstrated the responsiveness of intestinal strictures to biologics, specifically adalimumab, as evidenced by both clinical and radiological improvements [15, 16]. These findings suggest reconsidering the permanent assignment of disease behavior in CD, especially given evolving therapies and their impact on disease progression and management.

Classifying disease location, especially upper GIT involvement, led to notable disagreements among participants. Nearly 50% were divided on classification when VCE detected inflammation in the proximal and middle SB tertiles. While the IBD board recommended classifying these cases as upper GIT involvement, many participants indicated ileal disease with or without upper GIT involvement. This suggests that familiarity with VCE and its classification implications may be limited among gastroenterologists. The Montreal classification does not address VCE, but typically, inflammation in the first tertile suggests upper GIT involvement (L4). The SB3 VCE (Medtronic) divides the SB into three tertiles based on transit time [28], complicating the exclusion of proximal ileal disease, while the PillCam Crohn's

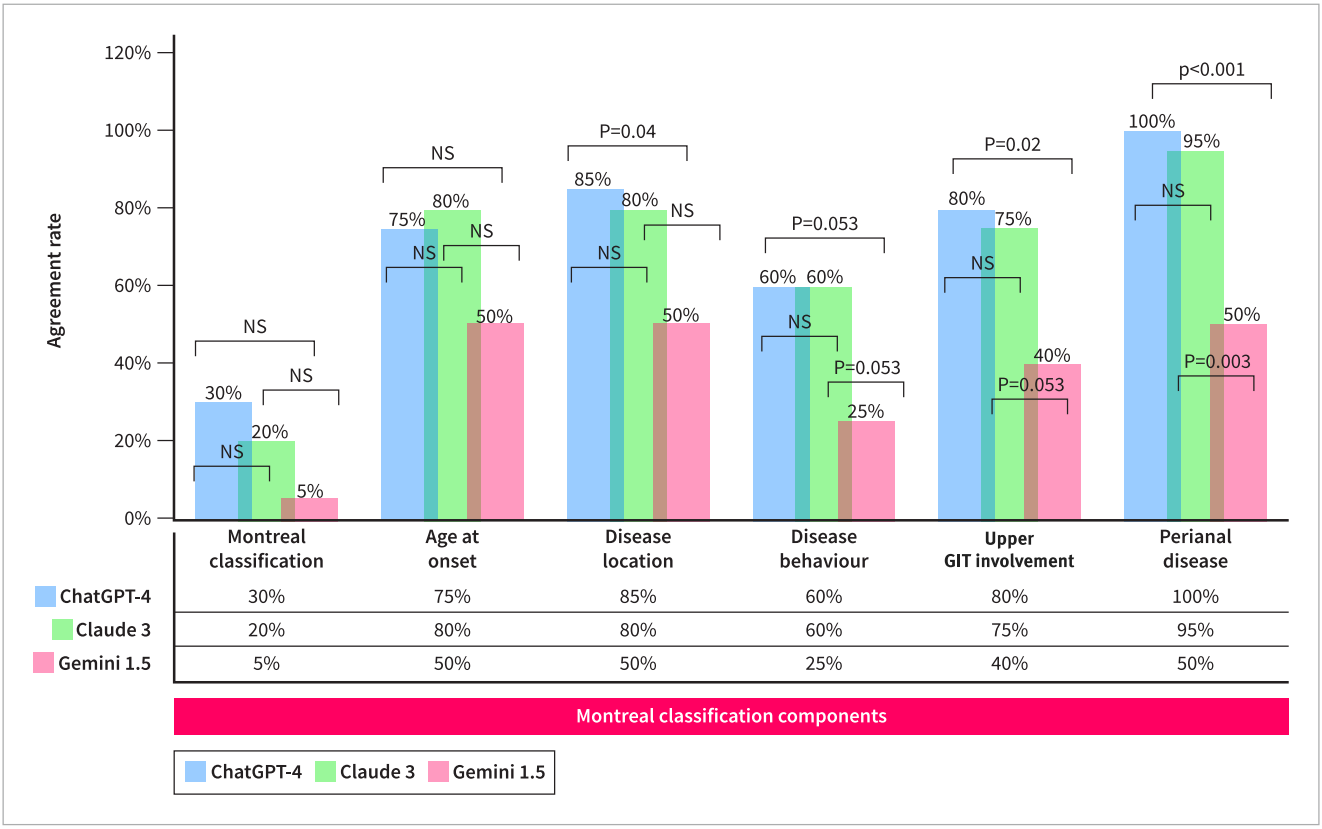

FIGURE 3 | Comparison of agreement rates among large language models with the inflammatory bowel disease board's consensus. GIT, gastrointestinal tract.

capsule (Medtronic) divides the bowel into three anatomical tertiles [29], potentially offering more precise location assignment, though its accuracy is not yet established.

The definition of upper GIT involvement has shifted from including any disease proximal to the terminal ileum (Vienna classification [10]) to excluding involvement distal to the jejunum (Montreal classification [9]). There was disagreement among participants regarding the classification of jejunal disease as L4, though the IBD board endorsed it as upper GIT involvement. This understanding needs reinforcement, as proximal disease is associated with a worse prognosis in CD patients [7]. It should be clarified that upper GIT involvement can coexist with ileo-(colonic) disease and L4 should be seen as a modifier for other disease locations [9]. Consistent with Dassopoulos et al. [19], the IBD board defined disease location based on the maximal cumulative extent throughout the disease course, despite previous recommendations to set it before surgery [9, 10]. Given advancements in diagnostic tools, re-evaluating this aspect of the Montreal classification could enhance its utility as a predictive tool.

The Montreal classification remains the most commonly used system for patients with CD. This classification provides a comprehensive assessment, including disease behavior, location, age, and perianal involvement, all of which have prognostic significance for patients with CD [3–5, 7]. Additionally, it was developed through an international consensus process, leading to widespread acceptance [8, 9]. Almost every study describing the baseline characteristics of CD patients employs components of the Montreal classification, which is crucial for standardization among researchers and readers. However, considering the findings of this study—particularly the gaps and disagreements among gastroenterologists in applying the Montreal classification, as well as advancements in diagnostic tools and biological therapies for CD [15, 16]—it is essential to refine and modify some definitions in the Montreal classification to improve patient management and maximize the benefits of this system.

Using three different LLMs to apply the Montreal classification revealed poor performance, especially in classifying disease behavior, compared with the IBD board. Comparing the LLMs as a whole may be misleading due to Gemini-1.5's inferior performance. Even when excluding Gemini-1.5 and including only ChatGPT-4 and Claude-3, their overall performance in the Montreal classification remained inferior to that of gastroenterologists. However, the LLMs performed better in classifying upper GIT and perianal involvement following Gemini-1.5 exclusion. While ChatGPT-4 and Claude-3 achieved success rates in some aspects of the Montreal classification similar to those reported for ChatGPT using the Truelove and Witts classification in ulcerative colitis (UC) patients (80%) [30], the complexity of the Montreal classification poses challenges for LLM accuracy, especially assigning disease phenotype. LLMs are generally accurate for straightforward questions based on structured guidelines, such as post-polypectomy surveillance and colorectal screening in IBD [31, 32]. However, as noted by Gravina et al., LLMs' performance is limited by insufficient

TABLE 3 | The notable points of disagreement in defining the Montreal classification of Crohn's disease among the study participants.

	Distribution of responses among participants	IBD board classification
Disease anatomic location (L)		
1. Inflammation of the proximal and middle SB tertiles as detected by VCE	Case 7: No classification—18 (47%) Ileal (L1)—20 (53%) Case 16: (this patient had also a colonic disease) Colonic (L2)—22 (58%) Ileo-colonic (L3)—15 (39%) No answer—1 (3%) Case 18: No classification—22 (58%) Ileal (L1)—16 (42%)	No classification
2. A history of past inflammation in the terminal ileum and current isolated inflammation of the colon	Case 19: Ileal (L1)—1 (3%) Colonic (L2)—12 (31%) Ileo-colonic (L3)—25 (66%)	L3
Disease behavior (B)		
1. Failure to intubate the ICV due to stenosis during ileo-colonoscopy while cross-sectional imaging shows no stenosis	Case 1: B1—11 (29%) B2—27 (71%) Case 5: B1—14 (37%) B2—22 (58%) Missing—2 (5%)	B1
2. Fistulizing perianal disease while no other penetrating features	Case 4: B1—28 (74%) B3—10 (26%)	Not enough to define B3
3. Previous features of obstructive or penetrating disease followed by resolution of the complicated features with or without immunosuppressive treatment.	Case 6: (history of SBO) B1—11 (29%) B2—27 (71%) Case 17: (History of entero-enteric fistula, maintaining remission following intestinal resection with fistulectomy. Recently presented with a stricture on MRE that resolved with immunosuppressive treatment)[a] B1—3 (8%) B2—2 (5%) B3—31 (82%) No answer—2 (5%) Case 20: (stricture by ileo-colonoscopy and MRE with resolution following immunosuppressive treatment)	B2/B3

(Continues)

TABLE 3 | (Continued)

	Distribution of responses among participants	IBD board classification
	B1—9 (24%) B2—27 (71%) No answer—2 (5%)	
4. Capsule retention without stricturing features upon cross-sectional imaging	Case 9: (PC retention with previously reported obstructive symptoms) B1—13 (34%) B2—24 (63%) No answer—1 (3%) Case 9: (VCE retention without obstructive symptoms) B1—20 (53%) B2—18 (47%)	B2 only if previously reported obstructive symptoms
5. Intestinal stenosis upon MRE without pre-stenotic dilation	Case 12: B1—9 (24%) B2—28 (73%) No answer—1 (3%)	B2
Upper GI tract involvement (L4)		
1. Inflammation of the proximal and middle SB tertiles as detected by VCE	Case 16: Upper GI (L4)—27 (71%) No—10 (26%) No answer—1 (3%)	L4
2. Inflammation of the jejunum	Case 12: Upper GI (L4)—25 (66%) No—13 (34%)	L4
3. History of an ileal-to-jejunal fistula in the distant past, maintaining remission for years, and now with terminal ileum disease involvement.	Case 12: Upper GI (L4)—19 (50%) No—19 (50%)	Not enough to define L4
4. Non specific chronic gastritis with negative *Helicobacter pylori* test and no history of nonsteroidal anti-inflammatory drug use	Case 15: Upper GI (L4)—25 (66%) No—13 (34%)	Not enough to define L4 unless having CD-related ulcers

Abbreviations: GI, gastrointestinal; IBD, inflammatory bowel disease; ICV, ileocecal valve; SB, small bowel; SBO, small bowel obstruction; VCE, video capsule endoscopy.
[a]Case 17 was added to this table because there were many double selections regarding disease behavior among the study participants, and we believed this case sparked considerable debate and controversy.

scientific references and outdated information [33], particularly in complex medical cases [17], which we believe is reflected in their application of the Montreal classification. Therefore, the use of LLMs cannot yet be endorsed for managing CD patients due to the complexity involved.

This study had several limitations. First, while we aimed to cover a broad range of CD presentations, ensuring full participation was challenging. Thus, the survey was designed to be practical and manageable. Nonetheless, the clinical vignettes included key aspects of CD and used diagnostic tools not previously evaluated for Montreal classification. Second, this survey included a relatively modest number of participants, despite our initial aim for a larger participant pool. Additionally, 87% of participants were practicing in tertiary centers. Nonetheless, this study represents the largest assessment of Montreal classification performance by gastroenterologists to date. Furthermore, the diverse geographical distribution of participants and their varying levels of expertise provided a comprehensive representation of CD management approaches. Finally, this study focused exclusively on CD, excluding UC cases, as CD's disease behavior and location present particular challenges for accurate classification.

In conclusion, this study highlights the challenges of accurately applying the Montreal classification to CD patients, particularly

concerning disease behavior and location, with a specific focus on the L4 assignment. We identified several points of notable disagreement in classifying CD patients, emphasizing the significant impact of newer diagnostic tools like VCE and the widespread use of others like MRE. This study underscores the need for improved education and revisions to some phenotype criteria to enhance consensus on the Montreal classification.

Author Contributions

O.U. and U.K. conceived and designed the study. O.U. developed the clinical cases, acquired and analyzed the data, and drafted the manuscript. UK contributed to drafting the manuscript. F.M., A.A., and H.Y. comprised the IBD board and participated in the study design. I.V., K.S. and S.B.-H. participated in study design. A.A., D.L., L.M., A.G., S.F., T. P., F.C., C.G.-M., J.L.-C., G.C., R.F.-I., N.T.-D., N.K.D., A.M., R.O., N.R., C.S.-C., G.D., S.R., A.F., L.C., T.B., A.E., S.L., C.F., V.M., I.C.R., J.J., I.B., H.T.D.S., L.B., M.B.D.A., P.R.A., A.D.B., Y.Z., X.S.-R., A.V. and U.K. participated in the survey. O.U., A.A., D.L., L.M., A.G., S.F., T.P., F.C., C.G.-M., J.L.-C., G.C., R.F.-I., N.T.-D., N.K.D., A.M., R.O., N.R., C.S.-C., G.D., S.R., A.F., L.C., T.B., A.E., S.L., C.F., V.M., I.C.R., J.J., I.B., H.T.D.S., L.B., M.B.D.A., P.R.A., A.D.B., Y.Z., X.S.-R., A.V., I.V., K.S., S.B.-H., H.Y., A.A., F.M. and U.K. participated in data interpretation and in critical revision of the manuscript for important intellectual property. All authors have approved the final draft submitted.

Conflicts of Interest

NKD received honoraria from Abbvie, Takeda and Cadigroup. CGM has received educational funding from Abbvie, Janssen, Kern Pharma, MSD, Tillots Pharma and has served as speaker for Galapagos. AA (Alessandro Armuzzi) received Consulting/advisory board fees from AbbVie, Alfa-Sigma, Amgen, Astra Zeneca, Biogen, Boehringer Ingelheim, Bristol-Myers Squibb, Celltrion, Eli-Lilly, Ferring, Galapagos, Gilead, Giuliani, Janssen, Lionhealth, Merck, Nestlé, Pfizer, Protagonist Therapeutics, Roche, Sanofi, Samsung Bioepis, Sandoz, Takeda, Tillots Pharma; Speaker's fees from AbbVie, AG Pharma, Amgen, Biogen, Bristol-Myers Squibb, Celltrion, Eli-Lilly, Ferring, Galapagos, Gilead, Janssen, Lionealth, Merck, Novartis, Pfizer, Roche, Samsung Bioepis, Sandoz, Takeda, Teva Pharmaceuticals. SBH has received advisory board and/or consulting fees from Abbvie, Takeda, Janssen, Celltrion, Pfizer, GSK, Ferring, Novartis, Roche, Gilead, NeoPharm, Predicta Med, Galmed, Medial Earlysign, BMS and Eli Lilly, holds stocks/options in Predicta Med, Evinature & Galmed, and received research support from Abbvie, Takeda, Janssen, Celltrion, Pfizer, & Galmed. UK received speaker and consultancy fees from Abbvie, BMS, Elly Lilly, Celtrion, Medtronic, Janssen and, Pfizer, Roche and Takeda, research support from Abbvie, Elli Lilly, Medtronic Takeda and Janssen. DL has received counseling, boards, transports or fees from Abbvie, Amgen, Biogaran, Biogen, Celltrion, Ferring, Galapagos, Janssen, Lilly, Medac, MSD, Pfizer, Prometheus, Sandoz, Takeda, Theradiag. HY received Advisory Committee or Review Panel fees from Takeda, Abbvie, Pfizer, Janssen, BMS, and Eli Lilli; Speaking and Teaching fees from Abbvie, Pfizer, Takeda, Novartis, and BMS; Grant/Research Support from Pfizer. NR has received lecture and consultancy fees from AbbVie, Janssen, and Takeda. FM received Lectures fees from AbbVie, Amgen, Biogen, Bristol Myers Squibb/Celgene, Celltrion, Dr Falk Foundation, Ferring Pharmaceuticals, Fresenius Kabi, Galapagos, Janssen, Lilly, MSD, Pfizer, Sandoz/Hexal, Takeda, Tillotts, and Vifor Pharma, Laboratorios Victoria. YZ have received support for conference attendance, speaker fees, research support and consulting fees from AbbVie, Adacyte Therapeutics, Almirall, Amgen, Dr. Falk Pharma, Faes Farma, Ferring Pharmaceuticals, Janssen, MSD, Otsuka, Pfizer, Shire, Takeda, Galapagos, Boehringer Ingelheim, Sanofi, Fresenius Kabi, Alfa-Sigma and Tillotts Pharma AG. MBA has served as a speaker, consultant and advisory member for or has received research funding from MSD, AbbVie, Janssen, Kern Pharma, Takeda, Galapagos-Alpha Sigma, Pfizer, Sandoz, Fresenius, Lilly, Ferring, Faes Farma, Dr. Falk Pharma, Chiesi, Adacyte and TillotsPharma. AA (Aurelien Amiot) received consulting fees from Abbvie, Pfizer, Takeda, Tillotts Pharma, Janssen and Sandoz as well as lecture fees and travel accommodations from Abbvie, Janssen, Pfizer, Takeda, Biogen, Fresenius Kabi, Amgen and Celltrion. LM has served as a speaker, consultant or advisory member for or has received unrestricted grants from MSD, Abbvie, Takeda, Janssen, Pfizer, Biogen, Galapagos, Kern Pharma, Lilly, Otsuka Pharmaceuticals, Tillotts, Dr. Falk Pharma, Ferring, Medtronic and General Electric. GD reports speaker fees and/or consultancy fees from Alfasigma, Celltrion Healthcare, Ferring, Johnson&Johnson, Novartis, Pfizer, and Takeda. HTS received speaker fees and travel accommodations from AbbVie, Biogen, Dr Falk Foundation, Ferring Pharmaceuticals, Janssen, Lilly, MSD, Pfizer, Takeda, Tillotts, and Laboratorios Victoria. The remaining authors declare no conflicts of interest.

Data Availability Statement

The data underlying this article will be shared on reasonable request by the corresponding authors.

References

1. M. Dolinger, J. Torres, and S. Vermeire, "Crohn's Disease," *Lancet (London, England)* 403, no. 10432 (2024): 1177–1191, https://doi.org/10.1016/S0140-6736(23)02586-2.

2. E. Louis, A. Collard, A. F. Oger, E. Degroote, F. A. Aboul Nasr El Yafi, and J. Belaiche, "Behaviour of Crohn's Disease According to the Vienna Classification: Changing Pattern Over the Course of the Disease," *Gut* 49, no. 6 (2001): 777–782, https://doi.org/10.1136/gut.49.6.777.

3. M. J. L. Romberg-Camps, P. C. Dagnelie, A. D. M. Kester, et al., "Influence of Phenotype at Diagnosis and of Other Potential Prognostic Factors on the Course of Inflammatory Bowel Disease," *American Journal of Gastroenterology* 104, no. 2 (2009): 371–383, https://doi.org/10.1038/ajg.2008.38.

4. A. V. Ramadas, S. Gunesh, G. A. O. Thomas, G. T. Williams, and A. B. Hawthorne, "Natural History of Crohn's Disease in a Population-Based Cohort From Cardiff (1986–2003): A Study of Changes in Medical Treatment and Surgical Resection Rates," *Gut* 59, no. 9 (2010): 1200–1206, https://doi.org/10.1136/gut.2009.202101.

5. I. C. Solberg, M. H. Vatn, O. Høie, et al., "Clinical Course in Crohn's Disease: Results of a Norwegian Population-Based Ten-Year Follow-Up Study," *Clinical Gastroenterology and Hepatology* 5, no. 12 (2007): 1430–1438, https://doi.org/10.1016/j.cgh.2007.09.002.

6. L. Peyrin-Biroulet, W. S. Harmsen, W. J. Tremaine, A. R. Zinsmeister, W. J. Sandborn, and E. V. J. Loftus, "Surgery in a Population-Based Cohort of Crohn's Disease From Olmsted County, Minnesota (1970–2004)," *American Journal of Gastroenterology* 107, no. 11 (2012): 1693–1701, https://doi.org/10.1038/ajg.2012.298.

7. M. Lazarev, C. Huang, A. Bitton, et al., "Relationship Between Proximal Crohn's Disease Location and Disease Behavior and Surgery: A Cross-Sectional Study of the IBD Genetics Consortium," *American Journal of Gastroenterology* 108, no. 1 (2013): 106–112, https://doi.org/10.1038/ajg.2012.389.

8. J. Satsangi, M. S. Silverberg, S. Vermeire, and J.-F. Colombel, "The Montreal Classification of Inflammatory Bowel Disease: Controversies, Consensus, and Implications," *Gut* 55, no. 6 (2006): 749–753, https://doi.org/10.1136/gut.2005.082909.

9. M. S. Silverberg, J. Satsangi, T. Ahmad, et al., "Toward an Integrated Clinical, Molecular and Serological Classification of Inflammatory Bowel Disease: Report of a Working Party of the 2005 Montreal World Congress of Gastroenterology,"supplement, *Canadian Journal of Gastroenterology* 19, no. SA (2005): 5A–36A, https://doi.org/10.1155/2005/269076.

10. C. Gasche, J. Scholmerich, J. Brynskov, et al., "A Simple Classification of Crohn's Disease: Report of the Working Party for the World

Congresses of Gastroenterology, Vienna 1998," *Inflammatory Bowel Diseases* 6, no. 1 (2000): 8–15, https://doi.org/10.1097/00054725-200002000-00002.

11. B. Bokemeyer, S. Plachta-Danielzik, R. di Giuseppe, et al., "Evaluation of a Downstaging, Bidirectional Version of the Montreal Classification of Crohn's Disease: Analysis of 5-Year Follow-Up Data From the Prospective BioCrohn Study," *Alimentary Pharmacology & Therapeutics* 58, no. 1 (2023): 35–47, https://doi.org/10.1111/apt.17512.

12. Y. Fan, L. Zhang, N. Omidakhsh, et al., "Patients With Stricturing or Penetrating Crohn's Disease Phenotypes Report High Disease Burden and Treatment Needs," *Inflammatory Bowel Diseases* 29, no. 6 (2023): 914–922, https://doi.org/10.1093/ibd/izac162.

13. G. Maconi, E. Bolzoni, A. Giussani, A. B. Friedman, and P. Duca, "Accuracy and Cost of Diagnostic Strategies for Patients With Suspected Crohn's Disease," *Journal of Crohn's and Colitis* 8, no. 12 (2014): 1684–1692, https://doi.org/10.1016/j.crohns.2014.08.005.

14. G. R. Lichtenstein, E. V. Loftus, K. L. Isaacs, M. D. Regueiro, L. B. Gerson, and B. E. Sands, "ACG Clinical Guideline: Management of Crohn's Disease in Adults," *Official Journal of the American College of Gastroenterology | ACG* 113, no. 4 (2018): 481–517, https://journals.lww.com/ajg/Fulltext/2018/04000/ACG_Clinical_Guideline__Management_of_Crohn_s.10.aspx.

15. J. D. Schulberg, E. K. Wright, B. A. Holt, et al., "Intensive Drug Therapy Versus Standard Drug Therapy for Symptomatic Intestinal Crohn's Disease Strictures (STRIDENT): An Open-Label, Single-Centre, Randomised Controlled Trial," *lancet Gastroenterol Hepatol* 7, no. 4 (2022): 318–331, https://doi.org/10.1016/S2468-1253(21)00393-9.

16. Y. Bouhnik, F. Carbonnel, D. Laharie, et al., "Efficacy of Adalimumab in Patients With Crohn's Disease and Symptomatic Small Bowel Stricture: A Multicentre, Prospective, Observational Cohort (CREOLE) Study," *Gut* 67, no. 1 (2018): 53–60, https://doi.org/10.1136/gutjnl-2016-312581.

17. Y.-J. Park, A. Pillai, J. Deng, et al., "Assessing the Research Landscape and Clinical Utility of Large Language Models: A Scoping Review," *BMC Medical Informatics and Decision Making* 24, no. 1 (2024): 72, https://doi.org/10.1186/s12911-024-02459-6.

18. A. J. Viera and J. M. Garrett, "Understanding Interobserver Agreement: The Kappa Statistic," *Family Medicine* 37, no. 5 (2005): 360–363.

19. T. Dassopoulos, G. C. Nguyen, A. Bitton, et al., "Assessment of Reliability and Validity of IBD Phenotyping Within the National Institutes of Diabetes and Digestive and Kidney Diseases (NIDDK) IBD Genetics Consortium (IBDGC)," *Inflammatory Bowel Diseases* 13, no. 8 (2007): 975–983, https://doi.org/10.1002/ibd.20144.

20. K. Krishnaprasad, J. M. Andrews, I. C. Lawrance, et al., "Inter-Observer Agreement for Crohn's Disease Sub-Phenotypes Using the Montreal Classification: How Good Are We? A Multi-Centre Australasian Study," *Journal of Crohn's and Colitis* 6, no. 3 (2012): 287–293, https://doi.org/10.1016/j.crohns.2011.08.016.

21. L. M. Spekhorst, M. C. Visschedijk, R. Alberts, et al., "Performance of the Montreal Classification for Inflammatory Bowel Diseases," *World Journal of Gastroenterology* 20, no. 41 (2014): 15374–15381, https://doi.org/10.3748/wjg.v20.i41.15374.

22. F. Rieder, D. Bettenworth, C. Ma, et al., "An Expert Consensus to Standardise Definitions, Diagnosis and Treatment Targets for Anti-Fibrotic Stricture Therapies in Crohn's Disease," *Alimentary Pharmacology & Therapeutics* 48, no. 3 (2018): 347–357, https://doi.org/10.1111/apt.14853.

23. D. Bettenworth, M. E. Baker, J. G. Fletcher, et al., "A Global Consensus on the Definitions, Diagnosis and Management of Fibrostenosing Small Bowel Crohn's Disease in Clinical Practice," *Nature Reviews Gastroenterology & Hepatology* 21, no. 8 (2024): 572–584, https://doi.org/10.1038/s41575-024-00935-y.

24. E. Rondonotti, "Capsule Retention: Prevention, Diagnosis and Management," *Annals of Translational Medicine* 5, no. 9 (2017): 198, https://doi.org/10.21037/atm.2017.03.15.

25. T. Sawada, M. Nakamura, O. Watanabe, et al., "Clinical Factors Related to False-Positive Rates of Patency Capsule Examination," *Therapeutic Advances in Gastroenterology* 10, no. 8 (2017): 589–598, https://doi.org/10.1177/1756283X17722744.

26. O. Ukashi, U. Kopylov, B. Ungar, et al., "Patency Capsule: A Novel Independent Predictor for Long-Term Outcomes Among Patients With Quiescent Crohn's Disease," *American Journal of Gastroenterology* 118, no. 6 (2023): 1019–1027, https://doi.org/10.14309/ajg.0000000000002118.

27. N. M. Noor, J. C. Lee, S. Bond, et al., "A Biomarker-Stratified Comparison of Top-Down versus Accelerated Step-Up Treatment Strategies for Patients With Newly Diagnosed Crohn's Disease (PROFILE): A Multicentre, Open-Label Randomised Controlled Trial," *lancet Gastroenterol Hepatol* 9, no. 5 (2024): 415–427, https://doi.org/10.1016/S2468-1253(24)00034-7.

28. I. M. Gralnek, R. Defranchis, E. Seidman, J. A. Leighton, P. Legnani, and B. S. Lewis, "Development of a Capsule Endoscopy Scoring Index for Small Bowel Mucosal Inflammatory Change," *Alimentary Pharmacology & Therapeutics* 27, no. 2 (2008): 146–154, https://doi.org/10.1111/j.1365-2036.2007.03556.x.

29. O. Ukashi, S. Soffer, E. Klang, R. Eliakim, S. Ben-Horin, and U. Kopylov, "Capsule Endoscopy in Inflammatory Bowel Disease: Panenteric Capsule Endoscopy and Application of Artificial Intelligence," *Gut and Liver* 17, no. 4 (2023): 516–528, https://doi.org/10.5009/gnl220507.

30. A. Levartovsky, S. Ben-Horin, U. Kopylov, E. Klang, and Y. Barash, "Towards AI-Augmented Clinical Decision-Making: An Examination of ChatGPT's Utility in Acute Ulcerative Colitis Presentations," *American Journal of Gastroenterology* 118, no. 12 (2023): 2283–2289, https://doi.org/10.14309/ajg.0000000000002483.

31. Y. Gorelik, I. Ghersin, I. Maza, and A. Klein, "Harnessing Language Models for Streamlined Postcolonoscopy Patient Management: A Novel Approach," *Gastrointestinal Endoscopy* 98, no. 4 (2023): 639–641.e4, https://doi.org/10.1016/j.gie.2023.06.025.

32. I. Ghersin, R. Weisshof, E. Koifman, E. Koifman, H. Bar-Yoseph, et al., "Comparative Evaluation of a Language Model and Human Specialists in the Application of European Guidelines for the Management of Inflammatory Bowel Diseases and Malignancies," *Endoscopy* 56, no. 09 (April 2024): 706–709, https://doi.org/10.1055/a-2289-5732.

33. A. G. Gravina, R. Pellegrino, M. Cipullo, et al., "May ChatGPT Be a Tool Producing Medical Information for Common Inflammatory Bowel Disease Patients' Questions? an Evidence-Controlled Analysis," *World Journal of Gastroenterology* 30, no. 1 (2024): 17–33, https://doi.org/10.3748/wjg.v30.i1.17.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.