









# Interobserver variability in colonoscopy quality assessment: a retrospective standardized multicenter video-based study

Radu Alexandru Vulpoi <sup>1</sup>, Tudor-Ştefan Rotaru <sup>1,2,\*</sup>, Mihaela Luca <sup>3</sup>, Adrian Ciobanu <sup>3</sup>, Cristian Gheorghe <sup>4</sup>, Eugen Dumitru <sup>5</sup>, Oana-Bogdana Bărboi <sup>1</sup>, Diana-Elena Floria <sup>1</sup>, Vadim Roşca <sup>1</sup>, Gheorghe Bălan <sup>1</sup>, Andrei Olteanu <sup>1</sup>, Vasile Liviu Drug <sup>1</sup>

<sup>1</sup>Department of Medical I, Discipline of Medical Semiology and Gastroenterology, Faculty of Medicine, Grigore T. Popa University of Medicine and Pharmacy, Iaşi, Romania. <sup>2</sup>Department of Medical Deontology and Bioethics, Faculty of Medicine, Grigore T. Popa University of Medicine and Pharmacy, Iaşi, Romania. <sup>3</sup>Institute of Computer Science, Romanian Academy, Iasi Branch, Iaşi, Romania. <sup>4</sup>Centre of Gastroenterology and Hepatology, Fundeni Clinical Institute, “Carol Davila” University of Medicine and Pharmacy, Bucharest, Romania. <sup>5</sup>Department of Gastroenterology, Faculty of Medicine, Ovidius University of Constanţa, Constanţa, Romania.

\*Correspondence: Tudor-Ştefan Rotaru, Grigore T. Popa University of Medicine and Pharmacy Iaşi, 16 Universitatii Str., Iaşi, Romania. Email: tudor.rotaru@umfiasi.ro

**How to cite this article:** Vulpoi RA, Rotaru TS, Luca M, et al. Interobserver variability in colonoscopy quality assessment: a retrospective standardized multicenter video-based study. *Arch Clin Cases*. 2026;13(2):30-37. doi: 10.22551/2026.51.1302.10338

## ABSTRACT

**Background and Aims:** Colonoscopy quality assessment is essential for adequate bowel preparation and complete examination, yet even validated tools such as the Boston Bowel Preparation Scale (BBPS) remain partly subjective. We assessed interobserver variability in expert evaluation using a standardized multicenter video dataset. **Methods:** This retrospective multicenter study included 64 anonymized complete colonoscopy videos from two academic centers. Eight experienced gastroenterologists independently evaluated recordings in randomly assigned pairs. Videos were assessed by five reviewer-pair combinations; individual reviewers evaluated between 10 and 33 examinations, and each pair assessed between 10 and 23 videos. Assessments included segmental and total BBPS scores, bowel preparation adequacy, and recognition of key anatomical landmarks: ileocecal valve, appendiceal orifice, hepatic and splenic flexures, and anal verge. Interobserver agreement was assessed using linear weighted Cohen's kappa for segmental BBPS scores, intraclass correlation coefficient (ICC) for total BBPS score, and Cohen's kappa with overall percent agreement for bowel preparation adequacy and anatomical landmark recognition. Because reviewer pairs varied across examinations, agreement measures were interpreted as pooled pairwise agreement across independent expert assessments. The Wilcoxon signed-rank test was retained as a complementary analysis for paired BBPS score differences. **Results:** Significant inter-reviewer variability was observed in BBPS scoring. Differences were found for the right colon, transverse colon, left colon, and total BBPS score: 2.42 vs 1.91,  $p < 0.01$ ; 2.47 vs 2.11,  $p < 0.01$ ; 2.44 vs 2.22,  $p < 0.05$ ; and 7.33 vs 6.23,  $p < 0.01$ , respectively. Overall, bowel preparation adequacy classification did not differ significantly, although discordant judgments occurred in 34% of examinations. Anatomical landmark recognition also varied, particularly for the appendiceal orifice and colonic flexures. **Conclusions:** Expert-based assessment may show clinically relevant variability despite standardized review conditions, supporting the need for more objective and reproducible approaches to colonoscopy quality control.

**KEYWORDS:** colonoscopy; bowel preparation; quality assessment; interobserver variability; Boston Bowel Preparation Scale; anatomical landmarks; cecal intubation

## 1. INTRODUCTION

Colonoscopy is considered the gold standard for the detection and prevention of colorectal neoplasia [1]. However, its effectiveness relies significantly on the quality of the procedure. Proper bowel preparation and a thorough examination are essential for accurate visual inspection of the

mucosa, effective lesion detection, and appropriate follow-up recommendations [2]. For this reason, quality assurance in colonoscopy has become a central concern in modern endoscopy practice [2–4].

Several validated indicators are currently used to assess the quality of colonoscopy [4–6]. The Boston Bowel Preparation Scale (BBPS) is commonly used to evaluate bowel preparation quality. It is a structured scoring system with proven validity and documented interobserver reliability, as shown in validation studies [7]. Similarly, confirming the

Received: April 2026; Accepted after review: May 2026;

Published: May 2026.



completeness of the examination usually involves documenting key anatomical landmarks, especially those that verify cecal intubation. In routine practice, however, even these standardized quality indicators remain partly dependent on human interpretation.

This dependence on expert assessment remains an important source of variability in the evaluation of colonoscopy quality [8]. Although scoring systems and predefined anatomical landmarks provide a standardized framework, they do not eliminate subjective interpretation. Assessment may still be influenced by the reviewer's clinical experience, level of attention, and personal opinion when assigning scores or confirming landmarks. Consequently, the same colonoscopy may be interpreted differently by different endoscopists, even when evaluated under standardized conditions and according to the same criteria.

This variability has practical clinical relevance. Differences in bowel preparation assessment can determine whether a colonoscopy is classified as adequate or inadequate, which may subsequently affect surveillance intervals, the need for repeat examination, and quality assessment metrics. Likewise, inconsistent identification of anatomical landmarks may influence the assessment of examination completeness, particularly cecal intubation [9]. Current quality guidelines place strong emphasis on adequate bowel preparation and proper documentation of complete colonoscopy as key performance indicators for both individual endoscopists and endoscopy units [10,11]. Therefore, inconsistent evaluation of these parameters may reduce the reliability of quality control in daily clinical practice.

Another consideration is that anatomical landmarks are not equally easy to identify. While some landmarks, such as the ileocecal valve or the anal verge, are generally recognized without difficulty, others are more difficult to confirm consistently, particularly when video documentation is brief, of suboptimal quality, or obtained without a stable view [12,13]. Previous studies have shown that reliable documentation of cecal intubation should ideally include at least two distinct cecal landmarks (the appendiceal orifice and the ileocecal valve) to improve consistency and reduce the risk of misclassification [2,4,14]. This highlights the importance not only of reaching the caecum but also of providing clear and reproducible endoscopic evidence of cecal intubation.

Recent work has explored automated approaches for bowel preparation assessment and other quality-related endoscopic tasks, suggesting that artificial intelligence may contribute to more consistent and scalable quality control [15–17]. However, before automated systems can be meaningfully developed and validated, it is necessary to define the extent and nature of variability in human assessment using standardized reference material.

Although previous studies have investigated interobserver variability in bowel preparation assessment or the recognition of specific anatomical landmarks, most have addressed these components separately or in less standardized settings [7,18–21]. The present study adds to this literature by assessing these components together in a standardized multicenter video-based dataset of complete colonoscopy examinations. Specifically, the study combined independent expert evaluation of segmental and total BBPS scores, bowel preparation adequacy classification, and recognition of key anatomical landmarks within the same assessment framework.

The primary objective of the present study was to quantify interobserver variability in colonoscopy quality assessment among experienced gastroenterologists reviewing standardized colonoscopy videos. We hypothesized that clinically relevant variability would still be observed even under standardized viewing and assessment conditions. By characterizing this variability across both bowel preparation scoring and landmark recognition, the study may help define areas in which complementary objective tools could support more reproducible colonoscopy quality assessment [8].

## ■ 2. MATERIALS AND METHODS

This retrospective, observational, multicenter study was based on full-length withdrawal-phase colonoscopy videos recorded during routine clinical practice. The study was designed to allow standardized reassessment of the same procedures by different reviewers under identical viewing conditions.

A total of 66 colonoscopy recordings were screened for eligibility. Eligible cases were complete colonoscopy videos from patients undergoing colorectal cancer screening at two tertiary academic centers in Romania. Screening-eligible patients included those at average risk, defined as individuals aged 50–74 years with a positive fecal immunochemical test (FIT), and those at increased risk due to a family history of colorectal cancer in a first-degree relative diagnosed after the age of 50 years [1,22,23]. Additional eligibility criteria included no personal history of colorectal cancer or colonic polyps, no prior colonic surgeries, and signed informed consent for colonoscopy. Videos had to demonstrate a complete examination of the caecum and meet quality standards for structured assessment. Exclusion criteria were incomplete exams, major technical issues, or insufficient data for evaluation. Two recordings from the Iași center were excluded due to technical problems. The final group consisted of 64 complete colonoscopy videos: 43 from the Institute of Gastroenterology and Hepatology at “Sf. Spiridon” Emergency University Hospital in Iași, and 21 from Fundeni Clinical Institute in Bucharest.

All procedures were digitally recorded during routine endoscopic practice using Olympus systems (EVIS EXERA III and EVIS X1), both of which provide high-definition imaging. To reduce technical heterogeneity, videos were captured using OBS (Open Broadcaster Software) with standardized settings for resolution, frame rate, and file format. The files were anonymized, assigned unique study codes, and stored in a dedicated study archive. File processing was limited to organization, integrity verification, and, when necessary, segmentation of relevant sequences for evaluation, without altering the clinical content of the recordings.

Two experienced gastroenterologists independently evaluated each of the 64 colonoscopy videos. In total, eight expert endoscopists participated in the review process, all with more than 10 years of experience in digestive endoscopy. Reviewers were blinded to each other's evaluations and received no additional clinical information during video review. Videos were assigned to five reviewer pairs, with each colonoscopy assessed by a single pair of reviewers. Each reviewer evaluated between 10 and 33 examinations, and each pair assessed between 10 and 23 videos. Some reviewers participated in more than one pair combination. The final reviewer allocation, including the number of videos

each reviewer assessed and the distribution of reviewer pairs, is provided in Supplementary Table 1. A standardized assessment form was used in all cases. No formal calibration session, pilot scoring exercise, or consensus discussion was performed before independent evaluation. All reviewers were experienced endoscopists familiar with BBPS scoring and standard colonoscopy quality indicators. This approach was chosen to reflect routine expert-based assessment under standardized review conditions, while preserving the independence of individual evaluations.

Bowel preparation quality was assessed using the Boston Bowel Preparation Scale (BBPS), with separate scores for the right, transverse, and left colon, according to validated methodology [7]. The total BBPS score was calculated as the sum of the three segmental scores. Bowel preparation was also classified as adequate or inadequate. Preparation was considered adequate when each colonic segment had a BBPS score of at least 2 and a total score of at least 6 [3,11,24]. The assessment form also included documentation of key anatomical landmarks relevant to complete examinations: the appendiceal orifice, ileocecal valve, hepatic flexure, splenic flexure, and anal verge [12,14,25]. For each landmark, reviewers selected one of three categories: identified, not identified, or uncertain. When a landmark was considered visible, the corresponding time interval in the video was also recorded. In addition, reviewers provided an overall assessment of whether the examination appeared balanced and uniform across colonic segments, as a general indicator of procedural quality.

The statistical analysis focused primarily on interobserver agreement and reproducibility, rather than only on paired differences between assessments. Because each colonoscopy was independently assessed by two experienced gastroenterologists, analyses were performed using the

colonoscopy as the unit of comparison. As reviewer pairs varied across examinations and some reviewers contributed to more than one pair, agreement statistics were interpreted as pooled pairwise agreement across independent expert assessments, rather than as agreement between two fixed raters.

Segmental BBPS scores were treated as ordinal variables, and interobserver agreement was quantified using linear weighted Cohen’s kappa. The total BBPS score was analyzed as a continuous summary score, and reproducibility was assessed using the intraclass correlation coefficient (ICC). Given the variable reviewer-pair structure, the ICC was used as a global estimate of reproducibility between paired independent expert assessments.

Bowel preparation adequacy was analyzed as a binary variable. Preparation was classified as adequate when each colonic segment had a BBPS score of at least 2 and the total BBPS score was at least 6. Agreement for adequacy classification was assessed using overall percent agreement and unweighted Cohen’s kappa. Discordant classifications were additionally reported using paired contingency tables.

Anatomical landmark recognition was analyzed as a categorical variable with three response options: identified, not identified, and uncertain. For each landmark, interobserver agreement was quantified using overall percent agreement and unweighted Cohen’s kappa, with the uncertain category retained as a separate category. The Wilcoxon signed-rank test was retained only as a secondary analysis to identify systematic directional differences between paired BBPS scores and was not interpreted as a measure of agreement. Paired BBPS score differences were summarized as mean paired differences, calculated as second assessment minus first assessment, with 95% confidence intervals. For Wilcoxon analyses, effect size was calculated as

**Table 1.** BBPS score comparison and interobserver agreement. **Panel A.** Paired comparison of BBPS scores.

BBPS segment/score	First assessment mean ± SD	Second assessment mean ± SD	Mean paired difference, second – first (95% CI)	Wilcoxon p-value	Effect size r
Right colon BBPS	1.91 ± 0.66	2.42 ± 0.61	0.52 (0.32 to 0.71)	<0.001	0.54
Transverse colon BBPS	2.11 ± 0.48	2.47 ± 0.62	0.36 (0.20 to 0.52)	<0.001	0.49
Left colon BBPS	2.22 ± 0.68	2.44 ± 0.69	0.22 (0.03 to 0.41)	0.028	0.27
Total BBPS score	6.23 ± 1.33	7.33 ± 1.61	1.09 (0.68 to 1.51)	<0.001	0.57

**Panel B.** Interobserver agreement for BBPS scores.

BBPS segment/score	Exact agreement, n (%)	Agreement statistic
Right colon BBPS	31/64 (48.4%)	weighted κ = 0.18
Transverse colon BBPS	32/64 (50.0%)	weighted κ = 0.19
Left colon BBPS	39/64 (60.9%)	weighted κ = 0.36
Total BBPS score	14/64 (21.9%)	ICC = 0.21

**Note:** Segmental BBPS agreement was assessed using linear weighted Cohen’s kappa. Total BBPS reproducibility was assessed using ICC as a global estimate of agreement between paired independent expert assessments. Mean paired differences are presented as second assessment minus first assessment, with 95% confidence intervals. The Wilcoxon signed-rank test was retained as a complementary analysis for paired score differences. Wilcoxon effect size was calculated as  $r = |Z|/\sqrt{N}$ , with  $N = 64$  paired assessments. As reviewer pairs varied across examinations, the first and second assessment columns represent paired assessment positions rather than two fixed raters.

$r = |Z|/\sqrt{N}$ , where Z represents the standardized Wilcoxon test statistic and N represent the number of paired observations. A p-value < 0.05 was considered statistically significant.

### 3. RESULTS

#### 3.1. Study sample

A total of 64 colonoscopy videos were included in the final analysis, generating 128 independent expert assessments. Video recordings were obtained from two tertiary centers, with 43 examinations from Iași and 21 from Bucharest.

#### 3.2. Bowel preparation assessment

Formal agreement analysis showed limited reproducibility of BBPS scoring between paired independent expert

assessments (Table 1, Panel B). Exact agreement for segmental BBPS scores was 31/64 examinations (48.4%) for the right colon, 32/64 (50.0%) for the transverse colon, and 39/64 (60.9%) for the left colon. Linear weighted Cohen’s kappa values were 0.18 for the right colon, 0.19 for the transverse colon, and 0.36 for the left colon, indicating limited agreement. For the total BBPS score, exact agreement was observed in 14/64 examinations (21.9%), and the ICC was 0.21, suggesting low reproducibility of the total score between independent expert assessments. According to commonly used interpretation frameworks, these values correspond to poor or low interobserver agreement.

As a complementary analysis, the Wilcoxon signed-rank test showed systematic paired differences in BBPS scoring (Table 1, Panel A). The second assessment position showed higher scores than the first assessment position for the right colon ( $2.42 \pm 0.61$  vs  $1.91 \pm 0.66$ ; mean paired difference 0.52, 95% CI 0.32 to 0.71;  $p < 0.001$ ), transverse colon ( $2.47 \pm 0.62$  vs  $2.11 \pm 0.48$ ; mean paired difference 0.36, 95% CI 0.20 to 0.52;  $p < 0.001$ ), left colon ( $2.44 \pm 0.69$  vs  $2.22 \pm 0.68$ ; mean paired difference 0.22, 95% CI 0.03 to 0.41;  $p = 0.028$ ), and total BBPS score ( $7.33 \pm 1.61$  vs  $6.23 \pm 1.33$ ; mean paired difference 1.09, 95% CI 0.68 to 1.51;  $p < 0.001$ ). These Wilcoxon analyses indicate directional differences between paired assessments, whereas interobserver agreement and reproducibility were quantified separately using weighted kappa and ICC statistics.

When bowel preparation was dichotomized as adequate versus inadequate, the two independent assessments were concordant in 42/64 examinations (65.6%) and discordant in 22/64 examinations (34.4%). Both assessments classified preparation as inadequate in 6 cases and adequate in 36 cases. In 17 cases, the first assessment classified preparation as inadequate, whereas the second as adequate; the reverse pattern occurred in 5 cases. Cohen’s kappa for adequacy classification was 0.16, indicating low agreement beyond chance. These results show that although most examinations received concordant adequacy classifications, clinically relevant discordance persisted in approximately one third of cases (Table 2).

### 3.3. Anatomical landmark recognition

Agreement for anatomical landmark recognition varied substantially across anatomical sites. For the appendiceal orifice, overall agreement was 47/64 examinations (73.4%), with a Cohen’s kappa of 0.46. The ileocecal valve showed high overall agreement, 58/64 examinations (90.6%), although Cohen’s kappa was lower ( $\kappa = 0.30$ ), reflecting the highly unbalanced distribution of responses, with most assessments classifying the landmark as identified.

Agreement was lower for colonic flexures. For the hepatic flexure, overall agreement was 25/64 examinations (39.1%), with a kappa value close to zero ( $\kappa \approx 0.00$ ). For the splenic

flexure, overall agreement was 23/64 examinations (35.9%), with a Cohen’s kappa of 0.11. These findings indicate poor reproducibility of flexure recognition between paired independent expert assessments.

For the anal verge, overall agreement was 56/64 examinations (87.5%), while Cohen’s kappa was -0.06. This apparent discrepancy between high raw agreement and low kappa is likely explained by the predominance of “identified” responses, which limits chance-corrected agreement estimates. This likely reflects the prevalence paradox of Cohen’s kappa, which may occur when category distributions are highly imbalanced. Overall, landmark recognition was more reproducible for the ileocecal valve and anal verge, whereas agreement was considerably lower for the colonic flexures (Table 3).

## 4. DISCUSSION

### 4.1. Main findings and interpretation

The present study demonstrates that colonoscopy quality assessment remains vulnerable to interobserver variability, even when performed by experienced endoscopists using standardized video material and a structured assessment protocol [20,26]. This was supported by formal agreement statistics, which showed limited reproducibility of BBPS scoring between paired independent expert assessments. Linear weighted Cohen’s kappa values were low for segmental BBPS scores, and the ICC for total BBPS score also indicated limited reproducibility. In addition, Wilcoxon analyses showed systematic directional differences between paired assessments across all segmental BBPS scores and the total BBPS score. Together, these findings indicate that the use of a validated scale does not eliminate subjectivity in bowel preparation assessment. Although previous validation studies have supported the reliability and validity of the BBPS [7], our findings suggest that its practical application may still be influenced by individual interpretation [27].

The second assessment position showed higher BBPS scores than the first assessment position across all colonic segments. Because reviewer pairs varied across examinations, this finding should not be interpreted as a fixed difference between two individual raters, but rather as evidence of variability in scoring thresholds across paired independent expert assessments. This pattern was particularly evident in the right colon, where bowel preparation may be more difficult to assess because residual liquid, foam, or partially removable debris can affect perceived mucosal visibility [28,29]. Although most examinations received concordant adequacy classifications, agreement beyond chance was low, with a Cohen’s kappa of 0.16. Indeed, 22/64 colonoscopies (34.4%) received discordant adequacy judgments, which could potentially influence post-procedural decisions and quality reporting [10,11].

**Table 2.** Agreement between paired independent assessments regarding bowel preparation adequacy.

		Reviewer nr. 2	
		inadequate bowel preparation**	adequate bowel preparation*
Reviewer nr. 1	inadequate bowel preparation**	6	17
	adequate bowel preparation*	5	36

\*Adequate bowel preparation – Total BBPS  $\geq 6$  and score per segment  $\geq 2$ .  
 \*\*Inadequate bowel preparation - Total BBPS  $< 6$  and/or score per segment  $< 2$ .  
 Overall agreement: 42/64 (65.6%); Discordance: 22/64 (34.4%); Cohen’s kappa: 0.16.

**Table 3.** Inter-reviewer variability in anatomical landmark classification.

Anatomical landmark	First assessment: identified / not identified/ uncertain	Second assessment: identified / not identified / uncertain	Overall agreement, n (%)	Cohen's κ
Appendiceal orifice	36 (56.3%) / 19 (29.7%) / 9 (14.1%)	52 (81.3%) / 11 (17.2%) / 1 (1.6%)	47/64 (73.4%)	0.46
Ileocecal valve	59 (92.2%) / 5 (7.8%) / 0 (0%)	60 (93.8%) / 1 (1.6%) / 3 (4.7%)	58/64 (90.6%)	0.30
Hepatic flexure	28 (43.8%) / 16 (25.0%) / 20 (31.3%)	42 (65.6%) / 6 (9.4%) / 16 (25.0%)	25/64 (39.1%)	0.00
Splenic flexure	16 (25.0%) / 32 (50.0%) / 16 (25.0%)	29 (45.3%) / 7 (10.9%) / 28 (43.8%)	23/64 (35.9%)	0.11
Anal verge	61 (95.3%) / 3 (4.7%) / 0 (0%)	59 (92.2%) / 4 (6.3%) / 1 (1.6%)	56/64 (87.5%)	-0.06

**Note:** Landmark recognition was analyzed as a three-category variable: identified, not identified, and uncertain. Cohen's kappa was calculated with the uncertain category retained as a separate response category. Percentages were calculated using the final dataset of 64 colonoscopy videos.

Interobserver variability was also evident in anatomical landmark recognition. Overall agreement was higher for the ileocecal valve and anal verge, but Cohen's kappa values were lower than raw agreement rates, reflecting the predominance of "identified" responses for these landmarks. The appendiceal orifice showed moderate agreement, whereas the colonic flexures showed greater inconsistency. For the hepatic and splenic flexures, both overall agreement and kappa values were low, indicating poor reproducibility of flexure recognition. These findings suggest that landmark recognition depends not only on the presence of the structure itself, but also on the quality, duration, and clarity of its visual documentation [21]. This is particularly important because current recommendations emphasize the need for reliable documentation of a complete colonoscopy, ideally by recognizing more than one cecal landmark [14]. The inclusion of an "uncertain" category in our assessment framework was useful in capturing this intermediate zone of expert interpretation rather than forcing a binary decision.

**4.2. Clinical relevance and implications**

The findings have practical implications for colonoscopy quality control. If experts reviewing the same standardized video material may reach different conclusions regarding bowel preparation adequacy or landmark identification, then routine quality assessment cannot be assumed to be fully consistent [28]. Formal agreement measures further support this interpretation by showing that variability was not limited to statistically significant paired differences, but was also reflected in limited agreement and reproducibility between independent expert assessments. This is relevant not only for clinical decision-making but also for audit processes, benchmarking, and performance monitoring at both the endoscopist and unit levels.

In this context, the study supports the need for more objective and reproducible approaches to quality assessment [16,17]. Artificial intelligence-based systems may play an important complementary role by reducing variability and increasing standardization, particularly for image-based tasks such as bowel preparation grading [15,30]. However, this perspective should be interpreted with caution, because AI systems are themselves developed, trained, and validated using datasets annotated by human experts. Therefore, variability in expert annotation may influence both the quality of training labels and the apparent performance of AI-based systems. This highlights the importance of using standardized annotation protocols, transparent reference

standards, and, when possible, multiple expert annotations or consensus-based labels when developing AI tools for colonoscopy quality assessment. The purpose of such tools is not to replace endoscopists, but to support them through more stable and transparent quality evaluation. The standardized multicenter video dataset developed in this study also represents an important foundation for future validation of such systems.

**4.3. Strengths and limitations**

The main strengths of this study lie in the use of a standardized multicenter video dataset, the independent evaluation by experienced gastroenterologists, and the structured assessment of both bowel preparation quality and anatomical landmark recognition [30]. Another strength of the study is the inclusion of formal agreement measures appropriate for ordinal, binary, and categorical variables, allowing interobserver variability to be assessed beyond paired score differences alone. Nevertheless, several limitations should be noted.

First, the retrospective design and relatively small sample size may reduce the external relevance of the findings. Second, although video-based assessment may improve consistency, it cannot fully capture the conditions of real-time colonoscopy, where dynamic maneuvers such as washing, suction, and patient repositioning may influence the perceived quality of the examination.

Third, as cases were assessed by different reviewer pairs, some of the observed variability may be attributable to pair-specific rating patterns. Because some reviewers participated in more than one pair combination and the number of videos assessed by individual reviewers was not identical, reviewer-specific scoring tendencies and clustering effects cannot be completely excluded. For this reason, agreement statistics should be interpreted as pooled pairwise agreement across independent expert assessments, rather than as agreement between two fixed raters.

Fourth, intraobserver reproducibility was not assessed, because reviewers did not repeat the evaluation of the same videos at a separate time point. Therefore, the present study cannot determine whether individual reviewers would have assigned the same BBPS scores or landmark classifications on repeated assessment.

Fifth, no formal calibration session, pilot scoring exercise, or consensus discussion was performed before formal evaluation. While this approach was intended to reflect routine expert-based assessment and preserve independent

interpretation, it may have contributed to variability in BBPS scoring and landmark recognition. Therefore, the observed variability should be interpreted in the context of uncalibrated, independent expert assessment under standardized viewing conditions.

Finally, because only expert reviewers were included, the degree of variability observed in this study may not reflect that seen in settings involving endoscopists with varying levels of experience. Overall, these findings show that expert-based assessment of colonoscopy quality may remain variable even under standardized review conditions. This supports the development of complementary objective tools to improve consistency in endoscopic quality control.

## 5. CONCLUSIONS

In conclusion, expert-based assessment of colonoscopy quality may show clinically relevant interobserver variability, even under standardized review conditions. In this study, limited agreement was observed for BBPS scoring, bowel preparation adequacy classification, and selected anatomical landmarks, particularly the colonic flexures. A clinically relevant proportion of examinations received discordant judgments regarding bowel preparation adequacy. These findings suggest that complementary objective tools may help improve consistency and standardization in colonoscopy quality assessment.

## Ethics approval and consent to participate

The study was conducted in accordance with the principles of the Declaration of Helsinki and applicable national and European regulations governing biomedical research and the protection of personal data. Ethical approval was obtained from the Research Ethics Committee of “Grigore T. Popa” University of Medicine and Pharmacy, Iași (Approval No. 150/13.02.2022), together with institutional approval from “Sf. Spiridon” County Emergency Clinical Hospital, Iași. All included patients had provided written informed consent for the use of their colonoscopy recordings in scientific research. Data processing complied with Regulation (EU) 2016/679 (GDPR). The study database contained only coded, non-identifiable information, and access to the files was restricted to members of the research team directly involved in data handling and analysis.

## Consent for publication

All participants provided informed consent to the use of anonymized colonoscopy video recordings for research and scientific publication. No identifiable personal data is included in this manuscript.

## Availability of data

The datasets generated and/or analyzed during the current study are not publicly available due to institutional and data protection restrictions. Still, they are available from the corresponding author on reasonable request and subject to approval by the participating institutions and applicable data protection regulations.

## Conflict of interest

The authors declare that they have no competing interests.

## Funding

This research received no external funding.

## Acknowledgments

The authors thank the medical teams from the participating centers for their support with data collection and video archiving, as well as all patients who consented to the use of their anonymized colonoscopy recordings for research purposes.

## REFERENCES

1. Kumar R, Lewis CR. Colon Cancer Screening. [Updated 2024 Sep 10]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. [Available from: <https://www.ncbi.nlm.nih.gov/books/NBK559064/> at 5/16/2026]
2. Rex DK. Key quality indicators in colonoscopy. *Gastroenterol Rep (Oxf)*. 2023 Mar 10;11:goad009. PMID: 36911141; PMCID: PMC10005623. doi: 10.1093/gastro/goad009.
3. Park SB, Cha JM. Quality indicators in colonoscopy: the chasm between ideal and reality. *Clin Endosc*. 2022 May;55(3):332-8. PMID: 35656625; PMCID: PMC9178135. doi: 10.5946/ce.2022.037.
4. Tiankanon K, Aniwon S. What are the priority quality indicators for colonoscopy in real-world clinical practice? *Dig Endosc*. 2024 Jan; 36(1):30-9. PMID: 37422906. doi: 10.1111/den.14635.
5. Kastenber D, Bertiger G, Brogadir S. Bowel preparation quality scales for colonoscopy. *World J Gastroenterol*. 2018 Jul 14;24(26): 2833-43. PMID: 30018478; PMCID: PMC6048432. doi: 10.3748/wjg.v24.i26.2833.
6. Shine R, Bui A, Burgess A. Quality indicators in colonoscopy: an evolving paradigm. *ANZ J Surg*. 2020 Mar;90(3):215-21. PMID: 32086869. doi: 10.1111/ans.15775.
7. Lai EJ, Calderwood AH, Doros G, et al. The Boston bowel preparation scale: a valid and reliable instrument for colonoscopy-oriented research. *Gastrointest Endosc*. 2009 Mar;69(3 Pt 2):620-5. PMID: 19136102; PMCID: PMC2763922. doi: 10.1016/j.gie.2008.05.057.
8. Lei II, Gaya DR, Robertson A, et al. Inter- and Intraobserver Variability in Bowel Preparation Scoring for Colon Capsule Endoscopy: Impact of AI-Assisted Assessment Feasibility Study. *Cancers (Basel)*. 2025 Aug 29;17(17):2840. PMID: 40940936; PMCID: PMC12427401. doi: 10.3390/cancers17172840.
9. Baker FA, Mari A, Nafrin S, et al. Predictors and colonoscopy outcomes of inadequate bowel cleansing: a 10-year experience in 28,725 patients. *Ann Gastroenterol*. 2019 Sep-Oct;32(5):457-62. PMID: 31474791; PMCID: PMC6686086. doi: 10.20524/aog.2019.0400.
10. Kaminski MF, Thomas-Gibson S, Bugajski M, et al. Performance measures for lower gastrointestinal endoscopy: a European Society of Gastrointestinal Endoscopy (ESGE) Quality Improvement Initiative. *Endoscopy*. 2017 Apr;49(4):378-97. PMID: 28268235. doi: 10.1055/s-0043-103411.
11. Hassan C, East J, Radaelli F, et al. Bowel preparation for colonoscopy: European Society of Gastrointestinal Endoscopy (ESGE) Guideline - Update 2019. *Endoscopy*. 2019 Aug;51(8):775-94. PMID: 31295746. doi: 10.1055/a-0959-0505.
12. Tang SJ, Wu R. Ileocecum: A Comprehensive Review. *Can J Gastroenterol Hepatol*. 2019 Feb 3;2019:1451835. PMID: 30854348; PMCID: PMC6378086. doi: 10.1155/2019/1451835.
13. Taghiakbari M, Hamidi Ghalehjegh S, Jehanno E, et al. Automated Detection of Anatomical Landmarks During Colonoscopy Using a Deep Learning Model. *J Can Assoc Gastroenterol*. 2023 May 2;6(4): 145-51. PMID: 37538187; PMCID: PMC10395661. doi: 10.1093/jcag/gwad017.
14. Moran B, Sehgal R, O'Morain N, et al. Impact of photodocumentation of caecal intubation on colonoscopy outcomes. *Irish Journal of Medical Science*. 2021 Nov;190(4):1397-402. PMID: 33471300. doi: 10.1007/s11845-020-02469-z.
15. Zhou W, Yao L, Wu H, et al. Multi-step validation of a deep learning-based system for the quantification of bowel preparation: a prospective, observational study. *Lancet Digit Health*. 2021 Nov;3(11): e697-e706. doi: 10.1016/S2589-7500(21)00109-6. Erratum in: *Lancet Digit Health*. 2021 Nov;3(11):e696. doi: 10.1016/S2589-7500(21)00237-5. PMID: 34538736.

16. Lee JY, Calderwood AH, Karnes W, et al. Artificial intelligence for the assessment of bowel preparation. *Gastrointest Endosc.* 2022 Mar; 95(3):512-18.e1. PMID: 34896100. doi: 10.1016/j.gie.2021.11.041.
17. Lee JY, Park J, Lee HJ, et al. Automatic assessment of bowel preparation by an artificial intelligence model and its clinical applicability. *J Gastroenterol Hepatol.* 2024 Sep;39(9):1917-23. PMID: 38766682. doi: 10.1111/jgh.16618.
18. Calderwood AH, Jacobson BC. Comprehensive validation of the Boston Bowel Preparation Scale. *Gastrointest Endosc.* 2010 Oct;72(4): 686-92. PMID: 20883845; PMCID: PMC2951305. doi: 10.1016/j.gie.2010.06.068.
19. Rostom A, Jolicoeur E. Validation of a new scale for the assessment of bowel preparation quality. *Gastrointest Endosc.* 2004 Apr; 59(4):482-6. PMID: 15044882. doi: 10.1016/s0016-5107(03)02875-x. Erratum in: *Gastrointest Endosc.* 2004 Aug;60(2):326.
20. Heron V, Parmar R, Ménard C, et al. Validating bowel preparation scales. *Endosc Int Open.* 2017 Dec;5(12):E1179-88. PMID: 29202001; PMCID: PMC5698009. doi: 10.1055/s-0043-119749.
21. Schelde-Olesen B, Bjørsum-Meyer T, Koulaouzidis A, et al. Interobserver agreement on landmark and flexure identification in colon capsule endoscopy. *Tech Coloproctol.* 2023 Dec;27(12):1219-25. PMID: 37036637; PMCID: PMC10638147. doi: 10.1007/s10151-023-02789-z.
22. Knudsen AB, Rutter CM, Peterse EFP, et al. Colorectal Cancer Screening: An Updated Modeling Study for the US Preventive Services Task Force. *JAMA.* 2021 May 18;325(19):1998-2011. PMID: 34003219; PMCID: PMC8409520. doi: 10.1001/jama.2021.5746.
23. Jayasinghe M, Prathiraja O, Caldera D, et al. Colon Cancer Screening Methods: 2023 Update. *Cureus.* 2023 Apr 12;15(4):e37509. PMID: 37193451; PMCID: PMC10182334. doi: 10.7759/cureus.37509.
24. Hsu WF, Chiu HM. Optimization of colonoscopy quality: Comprehensive review of the literature and future perspectives. *Dig Endosc.* 2023 Nov;35(7):822-34. PMID: 37381701. doi: 10.1111/den.14627.
25. Ahmad A, Saunders BP. Photodocumentation in colonoscopy: the need to do better? *Frontline Gastroenterol.* 2021 Aug 2;13(4):337-41. PMID: 35722601; PMCID: PMC9186039. doi: 10.1136/flgastro-2021-101903.
26. Heron V, Martel M, Bessissow T, et al. Comparison of the Boston Bowel Preparation Scale with an Auditable Application of the US Multi-Society Task Force Guidelines. *J Can Assoc Gastroenterol.* 2019 May;2(2):57-62. PMID: 31294366; PMCID: PMC6507282. doi: 10.1093/jcag/gwy027.
27. Hanzel J, Sey M, Ma C, et al. Existing Bowel Preparation Quality Scales Are Reliable in the Setting of Centralized Endoscopy Reading. *Dig Dis Sci.* 2023 Apr;68(4):1195-207. PMID: 36266592. doi: 10.1007/s10620-022-07729-9.
28. Lee HJ, Keum B, Cho YS, Cha JM. Interobserver Variation of Bowel Preparation for Colonoscopy. *Dig Dis Sci.* 2023 Nov;68(11):4140-7. PMID: 37740890. doi: 10.1007/s10620-023-08114-w.
29. Massinha P, Almeida N, Cunha I, Tomé L. Clinical Practice Impact of the Boston Bowel Preparation Scale in a European Country. *GE Port J Gastroenterol.* 2018 Sep;25(5):230-5. PMID: 30320161; PMCID: PMC6170922. doi: 10.1159/000485567.
30. Chen J, Wang G, Zhou J, et al. AI support for colonoscopy quality control using CNN and transformer architectures. *BMC Gastroenterol.* 2024 Aug 9;24(1):257. PMID: 39123140; PMCID: PMC11316311. doi: 10.1186/s12876-024-03354-0.

**Supplementary Table 1.** Reviewer allocation and pair distribution.**Panel A.** Reviewer-level allocation.

Reviewer ID	No. of videos assessed	Position as first assessment	Position as second assessment	Reviewer-pair combinations	Experience
R1	21	0	21	R1–R6: 10; R1–R7: 11	> 10 years
R2	33	33	0	R2–R5: 10; R2–R8: 23	> 10 years
R3	10	10	0	R3–R4: 10	> 10 years
R4	10	0	10	R3–R4: 10	> 10 years
R5	10	0	10	R2–R5: 10	> 10 years
R6	10	10	0	R1–R6: 10	> 10 years
R7	11	11	0	R1–R7: 11	> 10 years
R8	23	0	23	R2–R8: 23	> 10 years

**Panel B.** Reviewer-pair distribution.

Reviewer pair	No. of videos assessed	Video source	Case/video codes
R2–R5	10	la?i	1–10
R2–R8	23	la?i	11–20, 31–41, 44–45
R1–R6	10	la?i	21–30
R3–R4	10	Bucharest	FD1–FD10
R1–R7	11	Bucharest	FD11–FD21

**Note:** Reviewer IDs are anonymized. All reviewers were experienced gastroenterologists with more than 10 years of experience in digestive endoscopy. Each video was assessed independently by two reviewers. Because reviewer pairs varied across examinations and some reviewers participated in more than one pair combination, the agreement analysis was interpreted as pooled pairwise agreement across independent expert assessments rather than as agreement between two fixed raters.