Bowel cleansing quality evaluation in colon capsule endoscopy: what is the reference standard?

Benedicte Schelde-Olesen, Anastasios Koulaouzidis, Ulrik Deding, Ervin Toth, Konstantinos John Dabos, Abraham Eliakim, Cristina Carretero, Begoña González-Suárez, Xavier Dray, Thomas de Lange, Hanneke Beaumont, Emanuele Rondonotti, Uri Kopylov, Pierre Ellul, Enrique Pérez-Cuadrado-Robles, Alexander Robertson, Irene Stenfors, Alejandro Bojorquez, Stefania Piccirelli, Gitte Grunnet Raabe, Reuma Margalit-Yehuda, Isabel Barba, Giulia Scardino, Salome Ouazana and Thomas Bjørsum-Meyer

Abstract

Background: The diagnostic accuracy of colon capsule endoscopy (CCE) depends on a well-cleansed bowel. Evaluating the cleansing quality can be difficult with a substantial interobserver variation.

Objectives: Our primary aim was to establish a standard of agreement for bowel cleansing in CCE based on evaluations by expert readers. Then, we aimed to investigate the interobserver agreement on bowel cleansing.

Design: We conducted an interobserver agreement study on bowel cleansing quality.

Methods: Readers with different experience levels in CCE and colonoscopy evaluated bowel cleansing quality on the Leighton–Rex scale and Colon Capsule CLEansing Assessment and Report (CC-CLEAR), respectively. All evaluations were reported on an image level. A total of 24 readers rated 500 images on each scale.

Results: An expert opinion-based agreement standard could be set for poor and excellent cleansing but not for the spectrum in between, as the experts agreed on only a limited number of images representing fair and good cleansing. The overall interobserver agreement on the Leighton–Rex full scale was good (intraclass correlation coefficient (ICC) 0.84, 95% CI (0.82–0.85)) and remained good when stratified by experience level. On the full CC-CLEAR scale, the overall agreement was moderate (ICC 0.62, 95% CI (0.59–0.65)) and remained so when stratified by experience level.

Conclusion: The interobserver agreement was good for the Leighton–Rex scale and moderate for CC-CLEAR, irrespective of the reader's experience level. It was not possible to establish an expert-opinion standard of agreement for cleansing quality in CCE images. Dedicated training in using the scales may improve agreement and enable future algorithm calibration for artificial intelligence supported cleansing evaluation.

Trial registration: All included images were derived from the CAREforCOLON 2015 trial (Registered with The Regional Health Research Ethics Committee (Registration number: S-20190100), the Danish data protection agency (Ref. 19/29858), and ClinicalTrials.gov (registration number: NCT04049357)).

Keywords: artificial intelligence algorithm, bowel cleansing, colon capsule endoscopy, interobserver agreement

Received: 27 May 2024; revised manuscript accepted: 21 September 2024.

Ther Adv Gastroenterol

2024, Vol. 17: 1–11

DOI: 10.1177/17562848241290256

Correspondence to:
Benedicte Schelde-Olesen
Department of Surgery, Odense University Hospital, Svendborg, Baagoes Alle 31, Svendborg 5700, Denmark

Department of Clinical Research, University of Southern Denmark, Odense, Denmark
benedicte.schelde-olesen@rsyd.dk

Anastasios Koulaouzidis
Department of Clinical Research, University of Southern Denmark, Odense, Denmark

Department of Surgery, Odense University Hospital, Svendborg, Denmark

Department of Social Medicine and Public Health, Pomeranian Medical University, Szczecin, Poland

Ulrik Deding
Thomas Bjørsum-Meyer
Department of Clinical Research, University of Southern Denmark, Odense, Denmark

Department of Surgery, Odense University Hospital, Svendborg, Denmark

Ervin Toth
Department of Gastroenterology, Skåne University Hospital, Lund University, Malmö, Sweden

Konstantinos John Dabos
Department of Gastroenterology, St. John's Hospital, Livingston, Scotland, UK

Abraham Eliakim
Uri Kopylov
Reuma Margalit-Yehuda
Department of Gastroenterology, Sheba Medical Center, Tel Aviv, Israel

Cristina Carretero
Alejandro Bojorquez
Isabel Barba
Department of Gastroenterology, Clínica Universidad de Navarra, Pamplona, Spain

Begoña González-Suárez
Department of Gastroenterology, Endoscopy Unit, Hospital Clínic de Barceona, Barcelona, Spain

Xavier Dray
Salome Ouazana
Center for Digestive Endoscopy, Sorbonne University, Saint Antoine Hospital, APHP, Paris, France

Thomas de Lange
Department of Medicine and Emergencies, Sahlgrenska University Hospital, Västre Götalandsregionen, Sweden

Department of Molecular and Clinical Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

Hanneke Beaumont
Department of Gastroenterology & Hepatology, Amsterdam UMC, Amsterdam, The Netherlands

Emanuele Rondonotti
Giulia Scardino
Gastroenterology Unit, Valduce Hospital, Como, Italy

Pierre Ellul
Division of Gastroenterology, Mater Dei Hospital, Msida, Malta

Enrique Pérez-Cuadrado-Robles
Department of Gastroenterology, Georges-Pompidou European Hospital, Paris, France

Alexander Robertson
Department of Gastroenterology, Western General Hospital, Edinburgh, UK

Irene Stenfors
Department of Hereditary Cancer, Karolinska University Hospital, Stockholm, Sweden

Stefania Piccirelli
Department of Gastroenterology and Digestive Endoscopy, Fondazione Poliambulanza Istituto Ospedaliero, Brescia, Italy

Gitte Grunnet Raabe
Department of Surgery, Odense University Hospital, Svendborg, Denmark

Introduction

Colon capsule endoscopy's (CCE) diagnostic accuracy depends on the quality of bowel preparation. The evaluation of cleansing quality is instrumental in patient management, and a conservative grading strategy will lead to a high reinvestigation rate. In contrast, a lenient strategy can lead to missed pathology. The reference standard for cleansing quality evaluations is nonexistent, and readers are only supported in their decisions by different cleansing scales created for the purpose. Two such scales are the Leighton–Rex scale[1] and the Colon Capsule CLEansing Assessment and Report (CC-CLEAR).[2] The Leighton–Rex is a qualitative scale, whereas the CC-CLEAR is a quantitative scale based on the reader-estimated percentage of visualized mucosa. The interobserver agreement on the two scales has been investigated only in studies that have included a few readers.[1–3] Bowel cleansing is a topic discussed extensively; however, the optimal bowel preparation protocol has yet to be established. To carry out reliable comparisons of bowel cleansing regimens, we need to minimize the impact of reader variation on the evaluation of bowel cleansing. This requires a reference standard that we have not yet been able to set. Artificial intelligence (AI) algorithms have been presented as a possible solution to the inconsistency problem in bowel cleansing evaluations.[4,5]

This study aimed to create an agreement standard for bowel cleansing quality in CCE based on an assessment by a group of expert readers. This standard could serve as a baseline for calibrating an AI algorithm for automatic cleansing quality assessment. To ensure the quality of the agreement standard, we aimed to investigate the interobserver agreement on bowel cleansing quality in CCE readings and compare the level of agreement between the Leighton–Rex scale and CC-CLEAR.

Materials and methods

We conducted an interobserver agreement study on bowel cleansing quality evaluated on the Leighton–Rex scale and CC-CLEAR. The manuscript is prepared according to the STROBE guidelines.[6]

All included images were derived from the CAREforCOLON 2015 trial (Registered with The Regional Health Research Ethics Committee (Registration number: S-20190100, registration date: approval date: February 7, 2020), the Danish data protection agency (Ref. 19/29858), and ClinicalTrials.gov (registration number: NCT04049357, registration date: August 7, 2019, enrollment period: August 3, 2020 to December 16, 2022)).[7]

Bowel cleansing scales

The Leighton–Rex scale is a qualitative scoring system based on two parameters: a four-point cleansing level scale and a two-point bubbles effect scale.[1] Often, the bubbles effect grade is incorporated into the cleansing level to ease the practical use of the scale. We asked the readers to report on the overall cleansing combining the cleansing level and bubbles effect scale. We gave the readers the description of the four grades provided by Leighton and Rex. For grading of entire CCE investigations on this scale, the large bowel is divided into five segments (cecum, right colon, transverse colon, left colon, and rectum) and scored individually and overall. This segmentation was unnecessary as we based our study on images rather than videos.

CC-CLEAR is a quantitative scale.[2] The large bowel is divided into three segments (right colon, transverse colon, and left colon) and scored individually from 0 to 3 points depending on the percentage of visualized mucosa (<50% = 0 points, 50%–75% = 1 point, >75% = 2 points, >90% = 3 points). The sum of the scores for the individual segments determines the overall score. To use this scale, readers must be able to report the percentage of visualized mucosa, at least within the preset intervals. We, therefore, asked the readers to report a percentage within a single image from the large bowel.

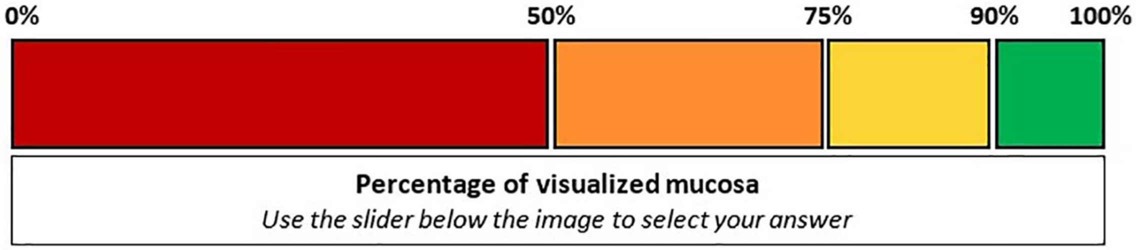
CCE readers

We invited endoscopists from several countries (Denmark, France, Israel, Italy, Malta, the Netherlands, Portugal, Scotland, Spain, and Sweden) with different levels of colonoscopy and CCE experience to participate in a CCE reader panel. We asked all invitees to place themselves in one of the following groups based on their level of experience:

- Group A: Readers extensively experienced in both colonoscopy (>5000 investigations) and CCE (>500 investigations).

Table 1. Instructions for capsule readers on the Leighton–Rex scale visible within the survey.

Excellent	Good	Fair	Poor
Not more than small bits of adherent fecal residue	Small amount of fecal residue, turbid fluid, bubbles (not interfering with examination)	Small amount of fecal residue, turbid fluid, or bubbles (partly interfering with examination)	Large amount of fecal residue or bubbles (which precludes a complete examination)



Percentage of visualized mucosa
Use the slider below the image to select your answer

Figure 1. Instructions for capsule readers on CC-CLEAR visible within the survey.
CC-CLEAR, Colon Capsule CLEansing Assessment and Report.

- Group B: Readers extensively experienced in colonoscopy (>5000 investigations) but without CCE experience.
- Group C: Readers slightly experienced in colonoscopy (250–1000 investigations) and CCE (10–50 investigations).
- Group D: Readers slightly experienced in colonoscopy (250–1000 investigations) but without CCE experience.

If readers did not fit into any group, we did not include them in the study. We included seven readers in groups A, B, and D, respectively. As it was difficult to find readers fitting the criteria for group C, only six were included in this group. All readers received a study protocol, including an introduction to both bowel cleansing scales.

CCE images and surveys

Through a review of a sample of CCE investigations carried out with the PillCam™ COLON2 capsules (Medtronic, Minneapolis, Minnesota, USA), we selected 500 CCE images and extracted them using the online platform PillCam Cloud Reader Software (Medtronic). We chose the images to represent the entire spectrum of bowel cleansing quality, so a randomized image selection could not be conducted. The endoscopist selecting the images was not part of the panel. The images were de-identified and presented in an online survey on a secure platform (SurveyXact; Rambøll, Aarhus, Denmark). We created two surveys, one for grading the cleansing quality on the Leighton–Rex scale and one for CC-CLEAR (percentage of visualized mucosa). The same 500 images were included in both surveys, meaning that the readers were presented with the same images twice. On the Leighton–Rex scale, readers were asked to report on a full scale marking one of four grades (*poor, fair, good, excellent*), and for the CC-CLEAR, a percentage of visualized mucosa (0%–100%) was selected using a sliding tool. A short instruction on the relevant grading scale was given within the survey and was always visible to the reader (Leighton–Rex: Table 1, CC-CLEAR: Figure 1). Half of the readers received a link to the Leighton–Rex and half to the CC-CLEAR survey. Readers were given 4 weeks to complete the task. We did a crossover 4 weeks after this deadline and distributed the second survey (Figure 2).

Agreement requirements

To create the standard of agreement for the calibration of the AI algorithm, we used evaluations by the experts in CCE and colonoscopy (group A) on the Leighton–Rex scale. For an image to be eligible for inclusion in the data material for the calibration, at least six out of seven experts should agree on the cleansing quality grade.

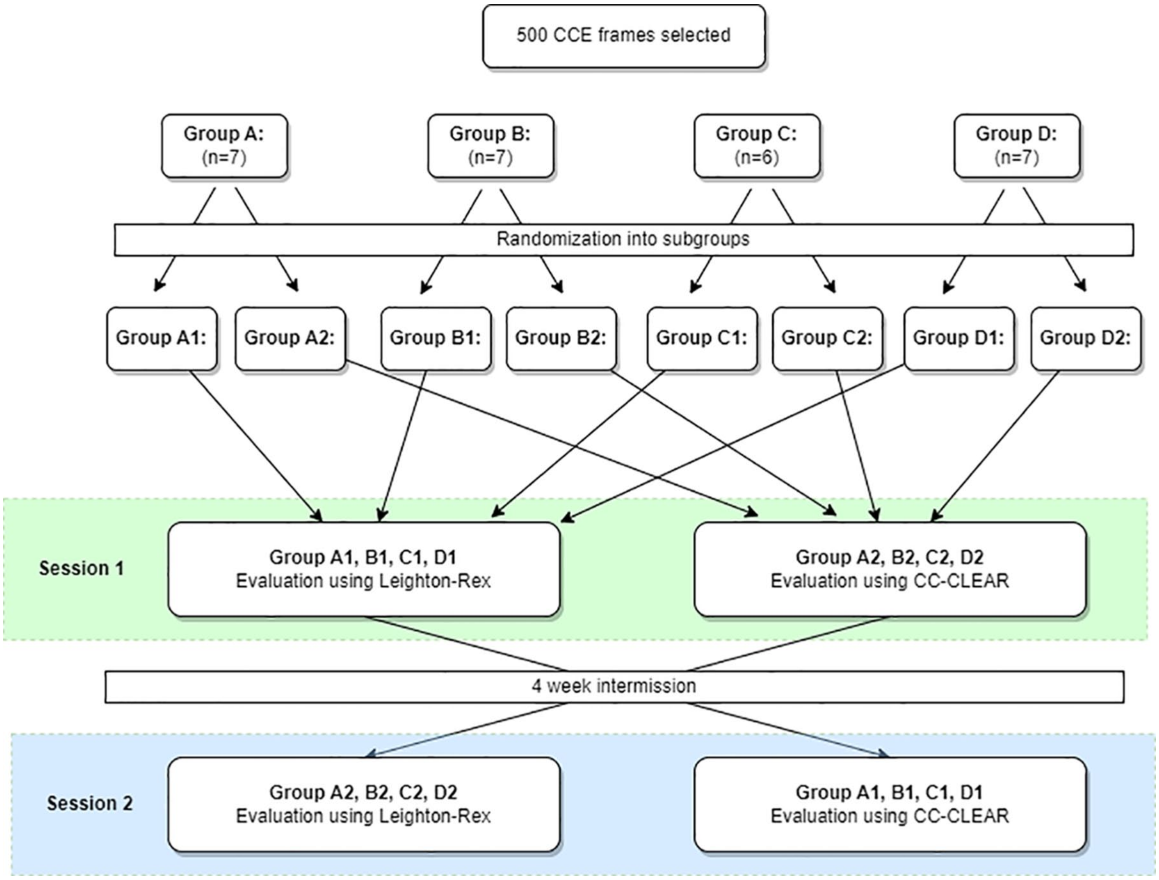

500 CCE frames selected
Group A: (n=7)
Group B: (n=7)
Group C: (n=6)
Group D: (n=7)
Randomization into subgroups
Group A1:
Group A2:
Group B1:
Group B2:
Group C1:
Group C2:
Group D1:
Group D2:
Session 1
Group A1, B1, C1, D1 Evaluation using Leighton-Rex
Group A2, B2, C2, D2 Evaluation using CC-CLEAR
4 week intermission
Session 2
Group A2, B2, C2, D2 Evaluation using Leighton-Rex
Group A1, B1, C1, D1 Evaluation using CC-CLEAR

Figure 2. Project setup and flow.

Sample size and statistical analysis

The sample size was based on the estimated number of images needed to calibrate the AI algorithm. We aimed for a minimum of 60 images per category (Leighton–Rex: poor, fair, good, excellent) with agreement between the desired number of experts. We included 500 images, more than double the necessary images for AI calibration to leave room for the likely variation.

We evaluated the interobserver agreement using the intraclass correlation coefficient (ICC). An ICC below 0.50 was interpreted as a poor agreement, 0.50–0.75 as a moderate agreement, 0.75–0.90 as a good agreement, and above 0.90 as an excellent agreement. Readers who did not complete both surveys were excluded from the analysis. A technical error occurred during the first session of the CC-CLEAR survey. This error was resolved, and readers were asked to redo the survey from the beginning if they encountered the problem. No issues occurred with the Leighton–Rex survey or in the second session of the CC-CLEAR survey. As the results for group B deviated substantially from the other groups, we performed subgroup analyses on subsets of images and readers in this group. First, one reader at a time was excluded from the analysis. Afterwards, we created subsets of images, analyzing them in groups as follows: 1–500, 101–500, 201–500, 301–500, and 401–500, to identify readers who had not restarted their CC-CLEAR evaluation following the survey error.

Results

Twenty-four readers completed both surveys (one incomplete in group C and two incompletes in group D).

Standard of agreement

We observed 80 images where 6 or 7 expert readers (group A) agreed on the Leighton–Rex grade *poor*, 21 for the grade fair, 27 for the grade good, and 122 for the grade *excellent*. A standard

Table 2. ICC for agreement between colon capsule endoscopy readers on the LR scale.

Group	LR full scale	LR simplified scale
	ICC (95% CI)	ICC (95% CI)
A, *n* = 7	0.85 (0.83–0.86)	0.73 (0.70–0.75)
B, *n* = 7	0.82 (0.80–0.84)	0.72 (0.69–0.74)
C, *n* = 5	0.83 (0.81–0.85)	0.74 (0.71–0.76)
D, *n* = 5	0.85 (0.83–0.87)	0.77 (0.75–0.80)
Overall, *n* = 24	0.84 (0.82–0.85)	0.73 (0.71–0.76)

ICC, intraclass correlation coefficient; LR, Leighton–Rex.

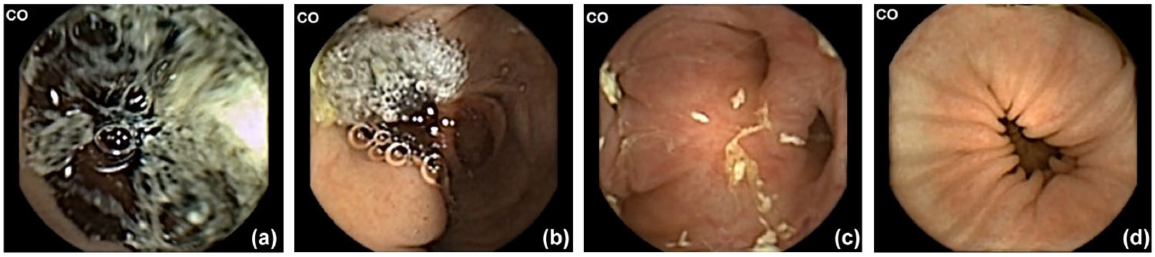


Figure 3. Examples of images with good agreement between readers on the Leighton–Rex scale. (a) Poor (24 readers in agreement). (b) Fair (21 readers in agreement). (c) Good (22 readers in agreement). (d) Excellent (24 readers in agreement).

of agreement for AI algorithm calibration was therefore not achieved (i.e. ⩾60 images with agreement) for grades *fair*, and *good*. If the number of desired readers in agreement was reduced to 5, the results were 103 images for *poor*, 51 for fair, 61 for good, and 139 for *excellent*.

Leighton–Rex, interobserver agreement

The overall agreement on the Leighton–Rex full scale (poor, fair, good, excellent) was good (ICC 0.84, 95% CI (0.82–0.85)). We observed the same level of agreement when stratifying by group/level of experience (Table 2). When considering the evaluations on a simplified scale (inadequate/adequate) where the grades poor/fair were deemed inadequate and good/excellent were considered adequate, the agreement declined to moderate both for the overall comparison and by group (Table 2). Examples of images with good agreement are depicted in Figure 3.

CC-CLEAR, interobserver agreement

When evaluating the reported percentages of visualized mucosa on a full scale (<50% = 0 points, 50%–75% = 1 point, >75% = 2 points, >90% = 3 points), the overall agreement was moderate (ICC 0.62, 95% CI (0.59–0.65)). Stratified by group, we also observed moderate agreement on the full scale (Table 3). When converting this to a simplified scale (inadequate/adequate) where a percentage of visualized mucosa below 75% was considered inadequate and above 75% was considered adequate, the overall agreement and the agreement for groups A, C, and D remained moderate. In contrast, the agreement for group B declined to poor. Finally, we looked at the percentage of visualized mucosa on a continuous scale from 0% to 100%. The overall agreement was poor (ICC 0.07, 95% CI (0.06–0.08)). This was mainly due to an inferior interobserver agreement in group B (ICC 0.02, 95% CI (−0.003, 0.04)), whereas the agreement in the other groups was moderate. We performed subgroup analyses to find a possible explanation for this low agreement within one group. First, on subsets of images, to test whether the poor agreement was caused by individuals not redoing the questionnaire after the error in survey distribution, as this could lead to better agreement in the subsets of images rated after the error was resolved. On

Table 3. ICC for agreement between colon capsule endoscopy readers on the CC-CLEAR scale.

Group	CC-CLEAR full scale	CC-CLEAR simplified scale	CC-CLEAR % of visualized mucosa
	ICC (95% CI)	ICC (95% CI)	ICC (95% CI)
A, *n* = 7	0.64 (0.60–0.67)	0.53 (0.50–0.57)	0.70 (0.67–0.73)
B, *n* = 7	0.56 (0.52–0.60)	0.45 (0.42–0.49)	0.02 (−0.003, 0.04)
C, *n* = 5	0.66 (0.63–0.70)	0.56 (0.52–0.60)	0.66 (0.63–0.69)
D, *n* = 5	0.66 (0.63–0.70)	0.55 (0.51–0.59)	0.70 (0.67–0.73)
Overall, *n* = 24	0.62 (0.59–0.65)	0.52 (0.49–0.55)	0.07 (0.06–0.08)
Excluding reader 13 (*n* = 23)	0.63 (0.60–0.66)	0.52 (0.49–0.56)	0.66 (0.63–0.69)

CC-CLEAR, Colon Capsule CLEansing Assessment and Report; ICC, intraclass correlation coefficient.

Table 4. Results for analysis on subsets of images for evaluation of the percentage of visualized mucosa in group B.

Image subset (group B)	CC-CLEAR % of visualized mucosa
	ICC (95% CI)
0–500	0.02 (−0.003, 0.04)
101–500	0.01 (−0.01, 0.04)
201–500	0.01 (−0.01, 0.04)
301–500	0.76 (0.72, 0.80)
401–500	0.80 (0.74, 0.85)

CC-CLEAR, Colon Capsule CLEansing Assessment and Report; ICC, intraclass correlation coefficient.

subsets of images, we found poor agreement in group B when including images 1–300 (Table 4). We observed good interobserver agreement when only including images 301–500. We then made subgroup analyses on all images, excluding one reader at a time, to test whether a single reader caused the error. When doing so, the interobserver agreement improved from poor to moderate when excluding reader 13 (Table 5). Excluding reader 13 from the overall analysis only affected the results significantly for CC-CLEAR as a continuous scale, where the interobserver agreement improved from poor to moderate (Table 3).

Discussion

The expert opinion-based standard of agreement was possible to set for poor and excellent bowel cleansing but not for the spectrum in between. We found that the agreement on the extremes, that is, images with large amounts of fecal matter, debris, etc., or images depicting a completely clean bowel, was much higher in the expert group than the images showing bowel cleansing of a more moderate quality. This is not surprising but leaves us with the problem of distinguishing between the cleansing grades that separate adequate from inadequate cleansing.

The overall interobserver agreement on the Leighton–Rex full scale was good and remained so when stratified by experience level. On the CC-CLEAR scale, divided into four grades, the interobserver agreement was moderate both overall and when stratified by experience level. We found a deviation in the results when looking at CC-CLEAR in group B separately, specifically the results for the evaluations on a continuous

Table 5. Results for analysis on subsets of readers for evaluation of the percentage of visualized mucosa in group B.

Excluded reader in group B	CC-CLEAR % of visualized mucosa
	ICC (95% CI)
Reader 8	0.01 (−0.01, 0.04)
Reader 9	0.01 (−0.01, 0.04)
Reader 10	0.01 (−0.01, 0.04)
Reader 11	0.01 (−0.01, 0.04)
Reader 12	0.02 (−0.01, 0.04)
Reader 13	0.61 (0.58, 0.65)
Reader 14	0.01 (−0.01, 0.04)

CC-CLEAR, Colon Capsule CLEansing Assessment and Report; ICC, intraclass correlation coefficient.

scale from 0% to 100%. A subgroup analysis showed that one reader in group B greatly affected the results. When this reader was excluded from the analysis, interobserver agreement improved from poor to moderate. The explanation for this could lie in a technical problem that may have caused a faulty response from one reader. The issue was corrected, which can explain why the interobserver agreement on the last 200 images in the dataset is much higher than what we see when including the entire dataset.

To sum up, we found moderate to good interobserver agreement. The agreement depended on the scale used but not the readers' experience level in CCE or colonoscopy. Readers with extensive colonoscopy experience and no experience in CCE (group B and D) could potentially tend to apply an evaluation strategy similar to that in colonoscopy. Still, the agreement within each group was similar.

In previous studies where good interobserver agreement was reported, a session where the readers evaluated CCE investigations in collaboration or reached a consensus on the bowel cleansing quality was carried out before the individual evaluations.[1,2] This practice will inherently calibrate the readers and lead to a unified understanding of the different bowel cleansing grades. We need the grading scales to be applicable in both a clinical and research setting and to enable us to compare results between readers, centers, and countries. The interobserver agreement should be studied without the calibration of readers since that is the reality we face. In a study by Buijs et al., where readers were not calibrated, the interobserver agreement on bowel cleansing quality was poor.[3] In our study, we presented the readers with a short description of the grading scales and referred them to the articles presenting the Leighton–Rex and CC-CLEAR scales. Even though we report a moderate to good interobserver agreement, the middle part of the cleansing quality spectrum still seems to pose a problem. On an image level, the cleansing quality scales do not lend the readers a strong enough support to yield a sufficient agreement. A short electronic introduction or a transferable teaching module between different settings could be a viable tool for creating a better background for using a specific scale and improving consensus.

As CCE is a younger modality on the endoscopy scene than small bowel (SB) capsule endoscopy, many issues related to capsule endoscopy have been discussed for years. Several cleansing quality grading scales have been proposed for SB capsule endoscopy, but no single scale has been accepted as the reference standard.[8–12]

Evaluating the cleansing quality is, unfortunately, not the only place where manual capsule reading faces obstacles. A recent meta-analysis on inter/intraobserver agreement in capsule endoscopy found that the agreement was suboptimal across various parameters in manual capsule endoscopy reading.[13] Often, AI is proposed as a solution to this issue.[14] By using AI, we can obtain a consistent result, minimizing the effect of observer variation. However, AI does need training based on inputs from human evaluations to establish the capabilities required to deliver a clinically relevant output. In this study, we aimed to create a standard of agreement for the cleansing quality in CCE, using the evaluations from an expert panel (group A). Several algorithms have been developed to classify bowel cleansing in SB capsule investigations[15–19] as well as in colonoscopy.[20–22] AI solutions specifically for CCE cleansing quality on a video level are still underway, but results on a machine learning approach have been published.[4] When comparing bowel cleansing quality evaluations by an AI to that of medical experts, the agreement is suboptimal in SB capsule endoscopy[23] and minimal in CCE.[24]

Cleansing quality is not the only parameter determining whether a CCE investigation can be considered adequate in terms of colon visualization. Such an evaluation requires assessment of the transit and possible technical interruptions. Furthermore, the information needed for further patient management includes a report on CCE findings and a characterization of those. The algorithm for cleansing quality evaluations is, therefore, only one piece of the puzzle in AI supported CCE reading. Several tasks are candidates for AI solutions such as cancer and polyp detection, size estimation of detected lesions, and localization of landmarks and findings. In the large AICE (AI-supported Image Analysis in Large Bowel Camera Capsule Endoscopy) project, a set of nine algorithms is being developed.[25] They are planned to work in collaboration to provide a thorough support for manual reading.

A strength of this study is the large number of capsule readers representing different nationalities and levels of experience in both capsule endoscopy reading and conventional endoscopy. This provides us with an impression of the challenges in using these grading scales in a real-world setting, where experience varies and doctors, as well as researchers, cross borders to carry out their work. We do acknowledge some limitations to our study. Because we carried out the study based on single CCE images instead of entire investigations, the results are not directly transferable to the experience in clinical practice. The scales we use are developed for videos and not for single images, which is a limitation to the setup of this study. However, in our opinion evaluating images can be seen as a necessary skill to master before assessing full CCE videos. One would assume that reaching a good agreement on images is much easier than videos. Considering this, the reported results do not underestimate the interobserver agreement on bowel cleansing quality, and the agreement might be lower if moving from single images to videos.

So the question remains: What is the reference standard for bowel cleansing quality in CCE? Further steps could involve expert group consensus where disagreements are resolved in a discussion on both CCE images and videos. Another approach could be to let the findings in CCE guide us toward a threshold for a cleansing quality that is adequate for detecting significant pathology. Coming to an agreement is paramount for elevating the quality of CCE and developing AI support for cleansing quality evaluations.

Conclusion

With input from an international selection of capsule readers and endoscopists, we found moderate to good interobserver agreement on bowel cleansing quality evaluations using the currently available cleansing scales for CCE. This agreement might, to some degree, be carried by agreement on the extremes in the spectrum of bowel cleansing and image-based evaluations as opposed to full-length investigations. The attempt by this study to create a standard of agreement for cleansing quality that we could utilize in training capsule readers and calibration of AI algorithms based on expert evaluations was not successful.

Declarations

Ethics approval and consent to participate

All included images were derived from the CAREforCOLON 2015 trial registered with The Regional Health Research Ethics Committee (Registration number: S-20190100, approval date: February 7, 2020). No separate approval for this specific study was required. No patients were included in this specific study. All patients included in the CAREforCOLON trial, from which the CCE images used in this study derived, gave informed consent for study participation.

Consent for publication

Not applicable.

Author contributions

Benedicte Schelde-Olesen: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Visualization; Writing – original draft; Writing – review & editing.

Anastasios Koulaouzidis: Conceptualization; Investigation; Methodology; Supervision; Writing – review & editing.

Ulrik Deding: Data curation; Formal analysis; Methodology; Visualization; Writing – review & editing.

Ervin Toth: Investigation; Writing – review & editing.

Konstantinos John Dabos: Investigation; Writing – review & editing.

Abraham Eliakim: Investigation; Writing – review & editing.

Cristina Carretero: Investigation; Writing – review & editing.

Begoña González-Suárez: Investigation; Writing – review & editing.

Xavier Dray: Investigation; Writing – review & editing.

Thomas de Lange: Investigation; Writing – review & editing.

Hanneke Beaumont: Investigation; Writing – review & editing.

Emanuele Rondonotti: Investigation; Writing – review & editing.

Uri Kopylov: Investigation; Writing – review & editing.

Pierre Ellul: Investigation; Writing – review & editing.

Enrique Pérez-Cuadrado-Robles: Investigation; Writing – review & editing.

Alexander Robertson: Investigation; Writing – review & editing.

Irene Stenfors: Investigation; Writing – review & editing.

Alejandro Bojorquez: Investigation; Writing – review & editing.

Stefania Piccirelli: Investigation; Writing – review & editing.

Gitte Grunnet Raabe: Investigation; Writing – review & editing.

Reuma Margalit-Yehuda: Investigation; Writing – review & editing.

Isabel Barba: Investigation; Writing – review & editing.

Giulia Scardino: Investigation; Writing – review & editing.

Salome Ouazana: Investigation; Writing – review & editing.

Thomas Bjørsum-Meyer: Conceptualization; Methodology; Supervision; Writing – review & editing.

Acknowledgements

We would like to acknowledge the extensive work of the CAREforCOLON study group.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The CAREforCOLON 2015 trial, from which the images used in this study originated, received funding from Aage and Johanne Louis-Hansen's Fond (grant 17–2B-1409), OUH's innovation fund (Grant R75-A3392), Medtronic Research Foundation (grant ERP 2018–11151), the Danish Cancer Society (grant R100-A6747), and the Excellence Centre in the Region of Southern Denmark (grant 18/48426). Funding was received from the University of Southern Denmark and Odense University Hospitals PhD fund.

Competing interests

Benedicte Schelde-Olesen has received honoraria and participated in advisory board meetings for Jinshan Ltd. Anastasios Koulaouzidis is shareholder of iCERV Ltd., has received consultancy fees and travel support from Jinshan Ltd., and has participated in advisory board meetings hosted by Dr FalkPharmaUK, Norgine, Jinshan, and ANKON. Emanuele Rondonotti has received honoraria from Fujifilm and consultancy fees from Medtronic. Xavier Dray has received lecture fees from Alfasigma, Bouchara, Recordati, Fujifilm, Medtronic, Norgine and Sandoz, has received consultancy fees from Norgine and Provepharma and is co-founder and shareholder of Augmented Endoscopy. Stefania Piccirelli has received a travel grant from AnX Robotica and honoraria from Malesci. Ulrik Deding, Ervin Toth, Konstantinos John Dabos, Abraham Eliakim, Cristina Carretero, Begoña González-Suárez, Thomas de Lange, Hanneke Beaumont, Uri Kopylov, Pierre Ellul, Enrique Pérez-Cuadrado-Robles, Alexander Robertson, Irene Stenfors, Alejandro Bojorquez, Gitte Grunnet Raabe, Reuma Margalit-Yehuda, Isabel Barba, Giulia Scardino, Salome Ouazana, Thomas Bjørsum-Meyer declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Availability of data and materials

Data will be made available upon reasonable request.

ORCID iDs

Benedicte Schelde-Olesen https://orcid.org/0000-0001-7643-2350

Anastasios Koulaouzidis https://orcid.org/0000-0002-2248-489X

Ervin Toth https://orcid.org/0000-0002-9314-9239

Thomas de Lange https://orcid.org/0000-0003-3989-7487

Uri Kopylov https://orcid.org/0000-0002-7156-0588

Reuma Margalit-Yehuda https://orcid.org/0000-0003-3910-7924

Supplemental material

Supplemental material for this article is available online.

References

1. Leighton JA and Rex DK. A grading scale to evaluate colon cleansing for the PillCam COLON capsule: a reliability study. *Endoscopy* 2011; 43: 123–127.
2. de Sousa Magalhães R, Arieira C, Boal Carvalho P, et al. Colon Capsule CLEansing Assessment and Report (CC-CLEAR): a new approach for evaluation of the quality of bowel preparation in capsule colonoscopy. *Gastrointest Endosc* 2021; 93: 212–223.
3. Buijs MM, Kroijer R, Kobaek-Larsen M, et al. Intra and inter-observer agreement on polyp detection in colon capsule endoscopy evaluations. *United European Gastroenterol J* 2018; 6: 1563–1568.
4. Buijs MM, Ramezani MH, Herp J, et al. Assessment of bowel cleansing quality in colon capsule endoscopy using machine learning: a pilot study. *Endosc Int Open* 2018; 6: E1044–E1050.
5. Becq A, Histace A, Camus M, et al. Development of a computed cleansing score to assess quality of bowel preparation in colon capsule endoscopy. *Endosc Int Open* 2018; 6: E844–E850.
6. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet* 2007; 370: 1453–1457.
7. Kaalby L, Deding U, Kobaek-Larsen M, et al. Colon capsule endoscopy in colorectal cancer screening: a randomised controlled trial. *BMJ Open Gastroenterol* 2020; 7: e000411.
8. Macedo Silva V, Lima Capela T, Freitas M, et al. Small Bowel CLEansing Assessment and Report (SB-CLEAR): standardizing bowel preparation report in capsule endoscopy. *J Gastroenterol Hepatol* 2023; 38: 747–751.
9. Alageeli M, Yan B, Alshankiti S, et al. KODA score: an updated and validated bowel preparation scale for patients undergoing small bowel capsule endoscopy. *Endosc Int Open* 2020; 8: E1011–E1017.
10. Brotz C, Nandi N, Conn M, et al. A validation study of 3 grading systems to evaluate small-bowel cleansing for wireless capsule endoscopy: a quantitative index, a qualitative evaluation, and an overall adequacy assessment. *Gastrointest Endosc* 2009; 69: 262–270, 270.e261.
11. Park SC, Keum B, Hyun JJ, et al. A novel cleansing score system for capsule endoscopy. *World J Gastroenterol* 2010; 16: 875–880.
12. Dray X, Houist G, Le Mouel JP, et al. Prospective evaluation of third-generation small bowel capsule endoscopy videos by independent readers demonstrates poor reproducibility of cleanliness classifications. *Clin Res Hepatol Gastroenterol* 2021; 45: 101612.
13. Cortegoso Valdivia P, Deding U, Bjørsum-Meyer T, et al. Inter/intra-observer agreement in video-capsule endoscopy: are we getting it all wrong? A systematic review and meta-analysis. *Diagnostics (Basel)* 2022; 12: 2400.
14. Moen S, Vuik FER, Kuipers EJ, et al. Artificial intelligence in colon capsule endoscopy—a systematic review. *Diagnostics (Basel)* 2022; 12(8): 1994.

15. Ribeiro T, Mascarenhas Saraiva MJ, Afonso J, et al. Design of a convolutional neural network as a deep learning tool for the automatic classification of small-bowel cleansing in capsule endoscopy. *Medicina (Kaunas)* 2023; 59(4): 810.

16. Nam JH, Hwang Y, Oh DJ, et al. Development of a deep learning-based software for calculating cleansing score in small bowel capsule endoscopy. *Sci Rep* 2021; 11: 4417.

17. Leenhardt R, Souchaud M, Houist G, et al. A neural network-based algorithm for assessing the cleanliness of small bowel during capsule endoscopy. *Endoscopy* 2021; 53: 932–936.

18. Noorda R, Nevárez A, Colomer A, et al. Automatic evaluation of degree of cleanliness in capsule endoscopy based on a novel CNN architecture. *Sci Rep* 2020; 10: 17706.

19. Klein A, Gizbar M, Bourke MJ, et al. Validated computed cleansing score for video capsule endoscopy. *Dig Endosc* 2016; 28: 564–569.

20. 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; 3: e697–e706.

21. Lee JY, Calderwood AH, Karnes W, et al. Artificial intelligence for the assessment of bowel preparation. *Gastrointest Endosc* 2022; 95: 512.e1–518.e1.

22. Zhou J, Wu L, Wan X, et al. A novel artificial intelligence system for the assessment of bowel preparation (with video). *Gastrointest Endosc* 2020; 91: 428.e2–435.e2.

23. Ju J, Oh HS, Lee YJ, et al. Clean mucosal area detection of gastroenterologists versus artificial intelligence in small bowel capsule endoscopy. *Medicine (Baltimore)* 2023; 102: e32883.

24. Schelde-Olesen B, Herp J, Braun J-M, et al. Interobserver agreement between an artificial intelligence algorithm and colon capsule endoscopy readers on bowel-cleansing quality. *iGIE* 2023; 2(2): 148.e3–153.e3.

25. AICE. AICE—simplifying the way we detect colon cancer with the AICE pathway, https://aiceproject.eu/ (2023, accessed June 27, 2024).