Thieme

Development and validation of a computed assessment of cleansing score for evaluation of quality of small-bowel visualization in capsule endoscopy

OPEN ACCESS

Authors

Einas Abou Ali[1], Aymeric Histace[2], Marine Camus[1,3], Rafaële Gerometta[4], Aymeric Becq[1,3], Olivia Pietri[1,3], Isabelle Nion-Larmurier[1], Cynthia Li[1,5], Ulriikka Chaput[1], Philippe Marteau[1,3], Christian Florent[1,3], Xavier Dray[1,2,3]

Institutions

1 Saint-Antoine Hospital, AP-HP, Department of Hepatogastroenterology, 184 rue du Faubourg Saint Antoine, 75012, Paris, France
2 ETIS UMR 8051, University Paris-Seine, University of Cergy-Pontoise, ENSEA, CNRS, Cergy, France
3 Sorbonne University, Paris, France
4 Saint Joseph Hospital, Paris, France
5 College of Arts and Sciences, Drexel University, Philadelphia, Pennsylvania, United States

submitted 12.8.2017
accepted after revision 25.1.2018

Bibliography
DOI https://doi.org/10.1055/a-0581-8758 |
Endoscopy International Open 2018; 06: E646–E651

ISSN 2364-3722

Corresponding author
Xavier Dray, MD, PhD, Endoscopy Unit, Sorbonne University & APHP Saint Antoine Hospital, 184 rue du Faubourg Saint Antoine, 75012 Paris, France
Fax: +0033-1-49-28-29-70
xavier.dray@aphp.fr

ABSTRACT

Background and study aims An objective and reliable scoring system is needed to assess quality of visualization in small bowel (SB) capsule endoscopy (CE), for both clinical practice and research purposes. The aim of this study was to establish and to validate a SB-computed assessment of cleansing (SB-CAC) score.

Patients and methods Thirty-three SB-CE were selected. A CAC score, defined as the ratio of the red over green pixels (R/G ratio), was calculated for each frame. Intervals were then determined, ranging from the lowest to the highest ratio among the extracted frames. Twelve frames were randomly selected in each of these intervals. Two hundred eighty-eight frames were shuffled and analyzed twice in random order by two experienced CE readers who were blinded to the CAC scores. Once an "adequately cleansed" or "inadequately cleansed" qualification was allotted to every still frame, a receiver operating characteristic (ROC) curve was created. In case of discrepancy between the two readers, the still frames were excluded. A second dataset of 288 different SB still frames was generated and read twice in random order by two other experienced SB-CE readers, using the same methodology.

Results A SB-CAC score threshold of 1.6 best achieved discrimination of adequately from inadequately cleansed frames, with a sensitivity of 92.7 % (95 %CI [89.7 – 95.8]) and a specificity of 92.9 % (95 %CI [89.9 – 95.9]). This threshold was validated using the second dataset, yielding the following performances: sensitivity 91.3 % (95 %CI [87.9 – 94.6]), specificity 94.7 % (95 %CI [92.1 – 97.3]).

Conclusion An SB-CAC score of 1.6 has the highest sensitivity and specificity to discriminate "adequately cleansed" from "inadequately cleansed" SB-CE still frames. This constitutes an objective, reproducible, reliable, and automated cleansing score for SB-CE.

Introduction

Quality of bowel preparation is of tremendous importance in capsule endoscopy (CE), as the device has no washing capability. Presence of bile, fluid, food residue, stools, and bubbles can impair visualization of mucosa, thus decreasing the diagnostic yield of CE examinations. While assessment of the quality of mucosal visualization during colonoscopy is strongly recommended [1, 2], there is no such recommendation in the setting of CE. Moreover, there is currently no consensus on the preparation regimen for small-bowel (SB) or colon CE. Most quantitative and qualitative scales used in trials to assess quality of visualization during CE are based on clinical evaluation and are not validated, with highly variable (poor to good) inter-observ-

er and intra-observer correlations [3,4]. Overall, we believe that there is a need for an objective, reliable, and reproducible scoring system to assess quality of visualization in CE, for both clinical practice and research. With that aim, Van Weyenberg et al. proposed a computed assessment of cleansing (CAC) score based on the ratio of color intensities of the red over green channel of the tissue color bar of CE video segments, to assess mucosal visibility [5]. The CAC score correlates well with other quantitative and qualitative scales. While a proof of principle was made with this score, there was no clear definition of adequate mucosal visualization according to the CAC score at the image level. Our aim was to develop and to validate an automated CAC score at the image level by defining the threshold of the red over green pixel ratio for an adequate SB visualization on still CE images.

Patients and methods

Patient and video selection

Inclusion criteria

Eligible participants presented with an indication for SB-CE for obscure gastro-intestinal bleeding (OGIB). SB-CE videos were de-identified and edited so that only the portion between the first image of the duodenum and the last image of the ileum were kept for analysis. The videos were then converted into mpeg files and included in the study.

Exclusion criteria

Patients were excluded if the SB-CE was a first-generation capsule (SB-CE1) or a third-generation capsule (SB-CE3, not available at the time of the study), if the procedure was incomplete, or if lesions of any kind were observed.

SB-CE2 procedure

Bowel preparation consisted of a clear liquid diet on the day prior to the procedure, followed by split ingestion of 1.5 L of polyethylene glycol-electrolyte (PEG) lavage solution: 1 L the evening before and 0.5 L on the morning of the procedure day. If the capsule was delayed in the stomach (over 1 hour), 10 mg metoclopramide could be administered orally. The procedure was complete when the capsule was expelled into the cecum. The second-generation capsule system used in this study consisted in the ingestible SB2 (PillCam®, Medtronic, Minnesota, United States).

Capsule image computerized analysis and selection

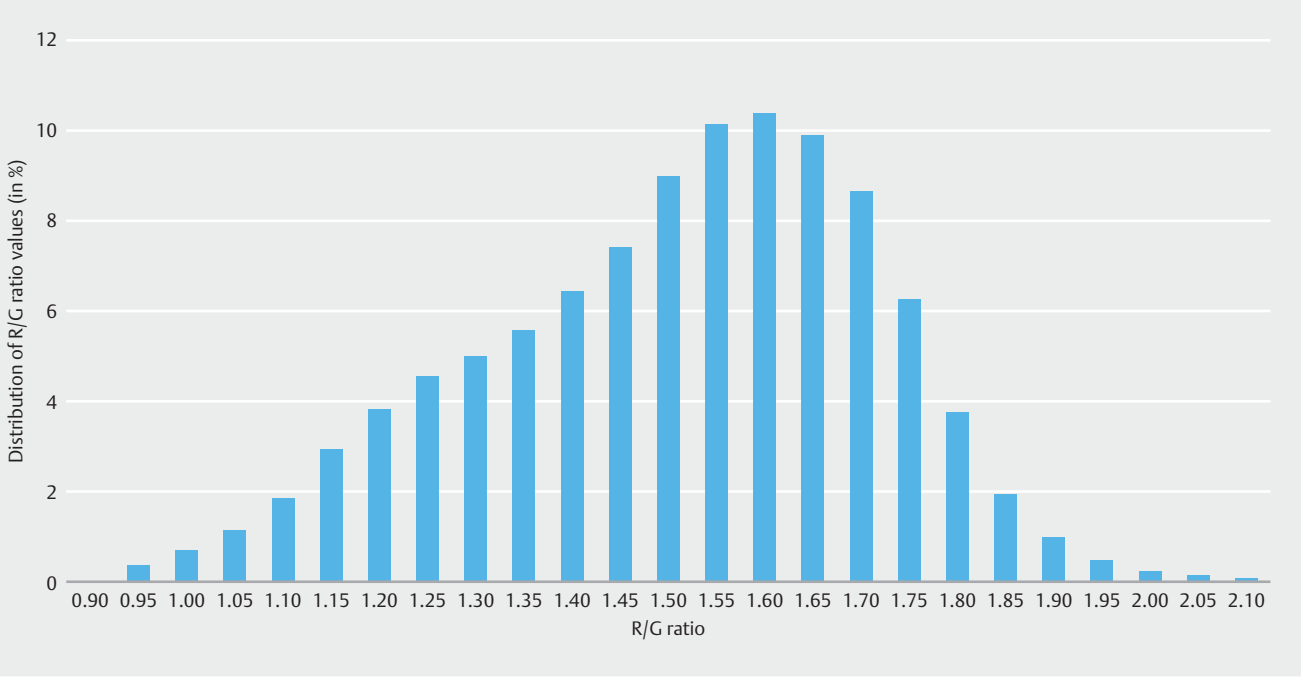
For each SB-CE2 video sequence, each still frame was individualized. The color intensities in the red (R), green (G) and blue (B) channels of each individual frame were extracted using MATLAB R2012a software (Mathworks, Natick, Massachusetts, United States). Our hypothesis, similar to that of Van Weyenberg et al. [5] at the tissue color bar level, was that a still frame of good quality of preparation is associated with high values of red intensity and low values of green intensity (higher R/G ratio), whereas a still frame of poor quality of preparation is associated with low values of red intensity and high values of green intensity (lower R/G ratio). Each still frame was allotted a CAC score based on red and green color intensities, formalized as R/G ratio. Subsequently, all still frames were sorted by ratio value to allow distribution assessment. The still frames were divided into groups composed of equal range, from the lowest to the highest CAC score. Afterwards, from all still frames from the 33 SB-CE2 procedures, 12 still frames were randomly selected from each interval (24 intervals of a 0.05 interval ranging from the lowest and highest ratio value). These frames constituted a panel representative of the variety of CAC score values (i. e. image quality). This procedure was repeated twice, once for both steps in the study (a development phase and then a validation phase). After sorting and random selection, two sets of still frames representative of the range of the R/G ratio were obtained for both steps of the study.

Capsule image expert review

Each set of still frames was analyzed by two experienced capsule readers who were blinded to the CAC values. The expert readers had previously analyzed over 500 SB-CE procedures. This review was performed twice for each set of still frames. Each set was shuffled between the two readings. Any still frame with visualization of over 90 % of the mucosa, with no, minimal or mild fluid and debris, bubbles, and bile/chyme staining, and with no, minimal or mild reduction of brightness was considered adequately cleansed according to the definition by Brotz et al. (▶ **Table 1**) [3]. Other frames were categorized as inadequate. Thus, each still frame was read four times (twice by both experts for each set). In the event of discrepancies (twice clas-

▶ **Table 1** Definition of "adequate" and "inadequate" cleansing of small bowel according to Brotz et al. [3].

Adequately cleansed	if excellent or good
Inadequately cleansed	if fair or poor
Excellent	Visualization of ≥ 90 % of mucosa; no, or minimal fluid and debris, bubbles, and bile/chyme staining; No, or minimal reduction of brightness.
Good	Visualization of ≥ 90 % of mucosa; mild fluid and debris, bubbles, and bile/chyme staining; Mildly reduced brightness.
Fair	Visualization of < 90 % of mucosa; moderate fluid and debris, bubbles, and bile/chyme staining; Moderately reduced brightness.
Poor	Visualization of < 80 % of mucosa; excessive fluid and debris, bubbles, and bile/chyme staining; Severely reduced brightness.



▶ **Fig. 1** Distribution of the computed assessment of cleansing (CAC) score (red over green ratio) among the 481,289 frames of 33 normal and complete small-bowel capsule endoscopy video sequences.

sified as adequate and twice as inadequate), still frames were excluded.

Distribution step: assessing proportion of frames with CAC scores over the cut-off value

To determine the distribution of the CAC score in clinical practice, a third set of 24 complete SB-CE videos was built and analyzed using the same methodology as the initial steps.

Statistics

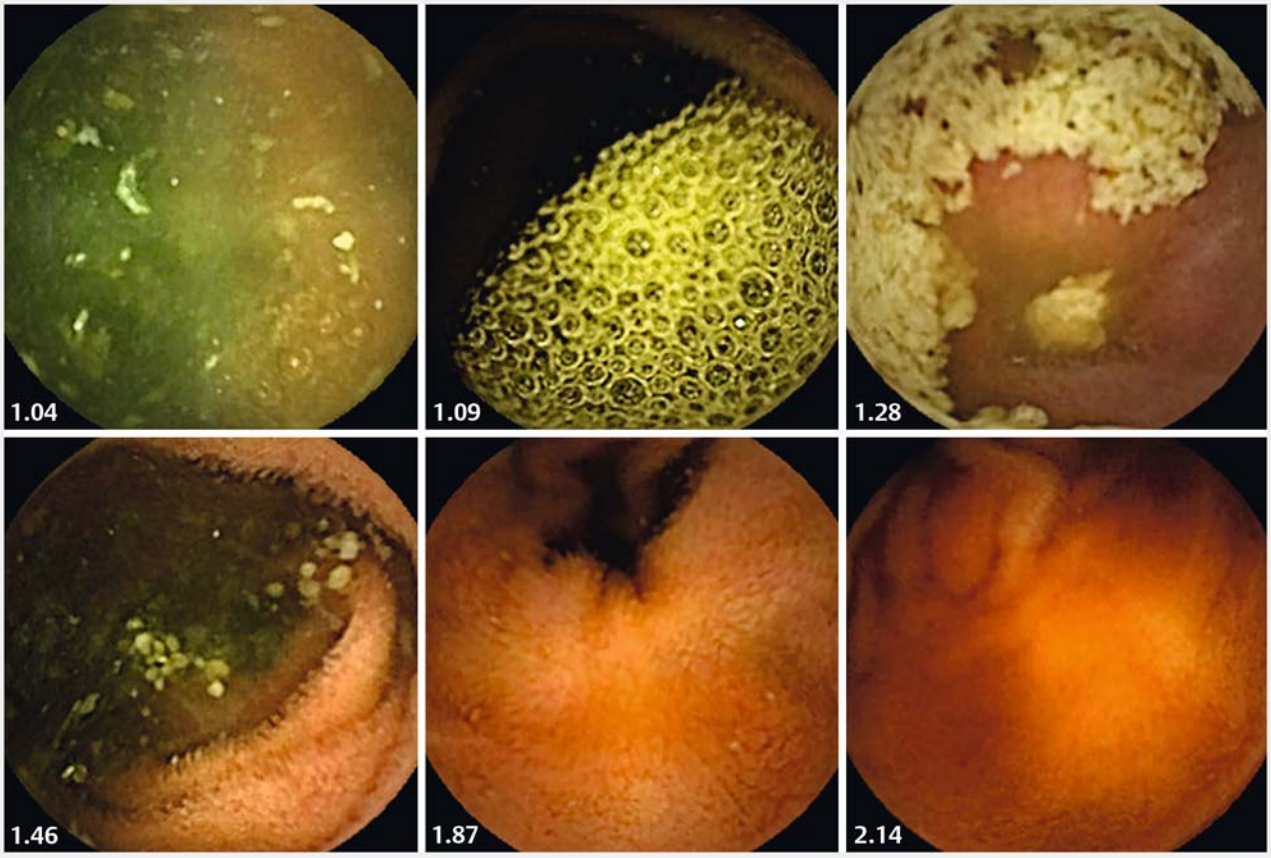
Quantitative variables were reported in mean and standard deviation values. Qualitative variables were reported in percentage values and 95 % confidence intervals (95 %CI). Pearson inter- and intra-observer correlation coefficients were also calculated. A receiver operating characteristic (ROC) curve was built using the R/G ratio intervals on the first set of still frames, with the expert reading as reference. By means of this ROC curve, a R/G ratio cut-off score was established, yielding the highest diagnostic performance in terms of discrimination between adequate and inadequate still frames with the highest operation point (i. e. trade-off between highest sensitivity and highest specificity). Sensitivity (Se, primary endpoint), specificity (Sp), positive (PPN) and negative predictive values (NPV, secondary end points) of the cut-off value were calculated on the first set of still frames (development phase). Then, this cut-off was validated using the second set of still frames (validation phase).

To assess the proportion of adequately cleansed still-frames using the validated CAC score (CAC score over the cut-off value) and their distribution (distribution phase), an ANOVA test was performed to assess the proportion and distribution of supposedly adequately cleansed frames (with CAC scores over the cut-off value) from the first to the fourth quartile using the third set of SB-CE videos [6]. The result of the ANOVA test was considered to be statistically significant for a P value < 0.05.

Results

Image selection and computed assessment of cleansing score feasibility

Between January 2014 and December 2015, 33 SB-CE procedures were included after first analysis, editing and de-identification. Fourteen (42.4 %) patients were men, and mean age was 64.5 years. Of these video sequences, 481,289 still frames were extracted. The color intensity in the red and green channel was measured for all still frames. The green intensity ranged from 37.55 to 86.09 (mean 63.67) and the red intensity ranged from 50.09 to 123.16 (mean 94.98). All color measurements were repeated and yielded the exact same results (Pearson intra-test correlation coefficient of 1.0). The R/G ratios ranged from 0.95005 to 2.1495 with a mean ratio of 1.4990. The distribution of the ratio values was determined, and based on the range between the lowest and highest ratio value, 24 intervals of a 0.05 interval value were created for the R/G ratios (▶ **Fig. 1**). ▶ **Fig. 2** shows SB-CE2 still frames representative of different CAC scores**.** Twelve still frames per interval were then randomly selected and included in the first and in the second sets. In total, 288 still frames were included in each individual set (the 288 still frames used in the second set are different from those of the firth set, but randomly selected with the same methodology).



▶ **Fig. 2** Examples of small-bowel frames with various computed assessment of cleansing (CAC) scores (= R/G ratio).

▶ **Table 2** Results of the four readings and experts' agreement of the two 228-frame datasets.

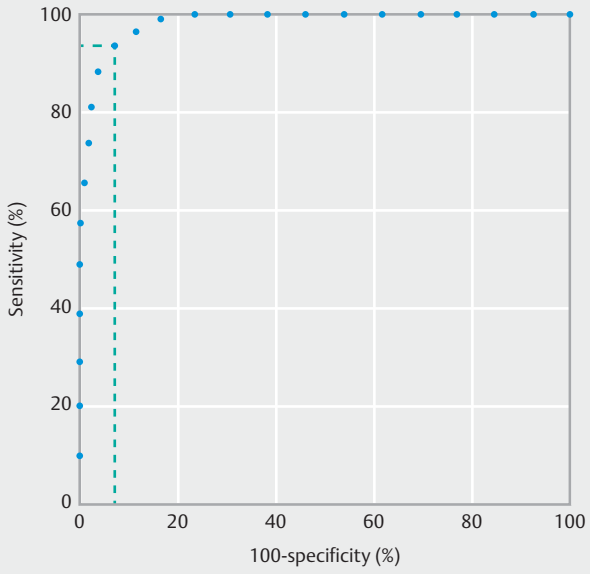
Dataset	Reader	Reading	Adequately cleansed frames (%)	Inadequately cleansed frames (%)	Pearson's coefficient	Exclusion of images for inter-reader discrepancies[1]
1	1	1	134 (46.5 %)	154 (53.5 %)	Intra-reader 1 = 0.90 Intra-reader 2 = 0.85 Inter-reader 1 – 2 = 0.87	8 frames
		2	126 (43.7 %)	162 (56.3 %)		
	2	1	121 (42.0 %)	167 (58.0 %)		
		2	126 (43.7 %)	162 (56.3 %)		
2	3	1	121 (42.0 %)	167 (58.0 %)	Intra-reader 3 = 0.89 Intra-reader 4 = 0.81 Inter-reader 3 – 4 = 0.82	11 frames
		2	119 (41.3 %)	169 (58.7 %)		
	4	1	104 (36.1 %)	184 (63.9 %)		
		2	110 (38.2 %)	178 (61.8 %)		

[1] Three or four agreements among the four readings of the same frame made a definitive classification. For any discrepancy (2 adequate and 2 inadequate classifications of the same frame), image was excluded from the analysis.

Still frame analysis

The expert reading results are described in ▶ **Table 2.**

For the first set of still images (development phase), 134 (46.5 %) and 126 (43.7 %) frames were classified as adequate, at first and second readings by reader 1, respectively. One-hundred-and-twenty-one (42.0 %) and 126 (43.7 %) still frames were classified as adequate, at first and second readings by reader 2, respectively. Pearson intra-observer correlation coefficients were 0.85 and 0.90. The Pearson inter-observer correla-



▶ **Fig. 3** Receiver operating characteristic (ROC) curve of the computed assessment of cleansing (CAC) score, according to expert classification of adequately or inadequately cleansed small bowel capsule endoscopy still frames.

tion coefficient was 0.87. Eight still frames were cases of discrepancies between the two readers and were therefore excluded. According to the ROC curve (▶ **Fig. 3**), a CAC score of 1.6 had the best performances to discriminate adequately from inadequately cleansed frames, with Se of 92.7% (95%CI [89.7–95.8]), Sp of 92.9% (95%CI [89.9–95.9]), PPV of 91.3 (95%CI [87.9–94.6]), and NPV of 94.2 (95%CI [91.4–96.9]).

For the second set of still images (validation phase), 121 (42.0%) and 119 (41.3%) were classified as adequate, at first and second readings by reader 3, respectively. One hundred and four (36.1%) and 110 (38.2%) still frames were classified as adequate at first and second readings, respectively, by reader 4. Pearson intra-observer correlation coefficients were 0.81 and 0.89. The Pearson inter-observer correlation coefficient was 0.82. Eleven still frames were cases of discrepancies between the two readers and were therefore excluded. The 1.6 CAC score cut-off was tested on this second dataset, giving the following performances: Se of 91.3% (95%CI [87.9–94.6]), Sp of 94.7% (95%CI [92.1–97.3]), PPV of 93.5 (95%CI [90.6–96,4]), and NPV of 92.9 (95%CI [89,82–95.9]).

The 24 SB-CE videos of the third set (distribution phase) contained a mean number of 13,262 ± 6,172 images. These videos were analyzed on a frame-by-frame basis for CAC scores in less than 1 hour. The mean (± SD) proportion of images with a CAC score over the 1.6 cut-off value (▶ **Table 3**) decreased significantly along the SB from 33.3 ± 23.3% in the first quartile, to 19.1 ± 22.5% in the second quartile, 20.8 ± 19.8% in the third quartile, and 17.3 ± 16.0% in the fourth quartile (ANOVA test, $P = 0.03$).

▶ **Table 3** Proportions (mean and standard deviations) of still frames with a computed assessment of cleansing (CAC) score (red over green pixel ratio) higher than the 1.6 cut-off ratio for each quartile (Q1 to Q4) of 24 normal and complete small-bowel capsule endoscopy video sequences.

	Q1 (%)	Q2 (%)	Q3 (%)	Q4 (%)
Mean	33.34	19.08	20.76	17.33
Standard deviation	23.28	22.46	19.85	15.98

Discussion

We validated a CAC scale, based on the red over green (R/G) pixel ratio of still frame images, to assess the quality of bowel cleansing in the setting of SB-CE. A SB-CAC score cut-off of 1.6 demonstrated a sensitivity of 91.3% and a specificity of 94.7%. The CAC score decreased significantly along the SB from the first to the fourth quartile. These results are consistent with what is already known in clinical practice: the proximal SB is usually cleaner than the distal ileum, probably due to accumulation of bile [6–8]. Moreover, our findings were consistent with those previously published by Van Weyenberg et al. [5], which also points to good external validity. In the study by Van Weyenberg et al., the authors made a proof of concept on use of a CAC score based on the red over green pixel ratio; there was a strong agreement between the computed scale and previously reported scales to assess the quality of small-bowel preparation. However, this score was based on the red over green pixel ratio of the tissue color bar. Examining the color tissue bar is a major loss of data compared to what is available from native images (as for diagnosis indeed): color is one thing, but other features (bubbles, luminosity, contrast, for instance) can be extracted from the native image (but not from the color tissue bar) and might contribute to improving diagnostic performance of the proposed algorithms. Our study provides more accurate data, as our SB-CAC is based on analysis of each individual SB still frame in a video segment. Moreover, we have determined a cut-off with a highly sensitive and specific threshold compared to expert reading, while no specific cut-off of the CAC score was given in the initial study by Van Weyenberg [5].

One strength of our study is that we built a solid foundation regarding the quality of bowel preparation to test the SB-CAC score. Still frames were randomly analyzed twice by experienced capsule readers who had no knowledge of the ratio values. A standardized and precise scale allowed reliable clinical assessment of the still frames quality of cleaning.

However, some limitations of this study must be acknowledged. First, still frames were evaluated rather than video sequences. In the future, entire SB-CE videos should be analyzed on a frame-by-frame basis and characteristics of an adequate video will have to be defined. Secondly, our sample of SB-CE videos was not representative of the general population as abnormal SB-CEs were excluded, and only cases of OGIB were chosen. Our main point is that any supposedly normal SB-CE should be reliable in terms of adequate preparation to be really conclusive in terms of normality. Thus, assessment of quality of

bowel preparation seems to be less important when an abnormality or an active bleeding is identified. We selected only cases of OGIB, because it is the most prevalent indication for SB-CE and because these patients were then representative of a homogenous population. Third, in the absence of preliminary data regarding development of such a R/G ratio at the image level, we were not able to estimate a number of frames needed to achieve sufficient statistical power. The sizes of image datasets were therefore arbitrarily chosen. However, the high diagnostic accuracy of the ratio, similar performances of the ratio in the two different datasets, and their narrow 95 % confidence intervals retrospectively suggest that the sample size was likely adequate. Finally, our study was performed when only second-generation SB-CE were available. The CAC-score needs to be validated with third-generation SB-CE.

Reporting on the quality of bowel preparation is important to render findings more reliable. It is firmly recommended for colonoscopy, where assessment tools (such as the Boston bowel preparation scale) are widely implemented in clinical practice [1, 9, 10]. However, there is no such recommendation for SB or colon CE. Quantitative and qualitative scales used in trials to assess quality of visualization on SB-CE videos sequences demonstrated inter- and intra-observer correlations ranging from 0.29 to 0.80, and from 0.45 to 0.76, respectively [4, 11]. However, when using a SB-CE qualitative scale for evaluation of still frames in the development and the validation phases of our study, we noticed inter- and intra-observer correlations higher than 0.80 (▶ **Table 2**). These findings support the idea that a frame-by-frame evaluation is more reliable and reproducible than an evaluation of a video sequence to assess quality of mucosal visualization. A full-length SB video sequence contains thousands of images with an important variation in quality of cleansing along the intestine, but a frame-by-frame evaluation of a full-length SB-CE video sequence is too demanding and very unlikely to ever be performed by a human reader. Nevertheless, it is feasible if any highly sensitive and specific computed algorithm is built.

Conclusion

In conclusion, the SB-CAC score based on the ratio of red over green pixels (R/G ratio) has a cut-off value of 1.6 with the highest sensitivity and specificity to discriminate "adequately cleansed" from "inadequately cleansed" SB-CE still frames. This score constitutes an objective, reproducible, reliable, automated, fast, and comprehensive cleansing score for SB-CE, and circumvents the subjectivity of qualitative grading systems. These findings set a path for future studies assessing the proportion of "adequately cleansed" frames in a SB-CE, allowing comparison of different bowel-cleansing regimens. Further research is warranted to determine which proportion of "adequately cleansed" frames defines an acceptable quality of preparation of SB-CE in clinical practice.

Competing interests

Xavier Dray has acted as a consultant for Boston Scientific, Fujifilm, Pentax, and Medtronic.

References

[1] 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; 69: 620 – 625

[2] Lieberman D, Nadel M, Smith RA et al. Standardized colonoscopy reporting and data system: report of the Quality Assurance Task Group of the National Colorectal Cancer Roundtable. Gastrointest Endosc 2007; 65: 757 – 766

[3] 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

[4] Goyal J, Goel A, McGwin G et al. Analysis of a grading system to assess the quality of small-bowel preparation for capsule endoscopy: in search of the Holy Grail. Endosc Int Open 2014; 2: E183 – E186

[5] Van Weyenberg SJB, De Leest HTJI, Mulder CJJ. Description of a novel grading system to assess the quality of bowel preparation in video capsule endoscopy. Endoscopy 2011; 43: 406 – 411

[6] van Tuyl SA, den Ouden H, Stolk MF et al. Optimal preparation for video capsule endoscopy: a prospective, randomized, single-blind study. Endoscopy 2007; 39: 1037 – 1040

[7] Dai N, Gubler C, Hengstler P et al. Improved capsule endoscopy after bowel preparation. Gastrointest Endosc 2005; 61: 28 – 31

[8] Ben-Soussan E, Savoye G, Antonietti M et al. Is a 2-liter PEG preparation useful before capsule endoscopy? J Clin Gastroenterol 2005; 39: 381 – 384

[9] Rees CJ, Bevan R, Zimmermann-Fraedrich K et al. Expert opinions and scientific evidence for colonoscopy key performance indicators. Gut 2016; 12: 2045 – 2060

[10] Calderwood AH, Jacobson BC. Comprehensive validation of the Boston Bowel Preparation Scale. Gastrointest Endosc 2010; 72: 686 – 692

[11] Park SC, Keum B, Hyun JJ et al. A novel cleansing score system for capsule endoscopy. World J Gastroenterol 2010; 16: 875 – 880