Multi-criterion, automated, high-performance, rapid tool for assessing mucosal visualization quality of still images in small bowel capsule endoscopy

OPEN ACCESS

Authors
Sarra Oumrani[1], Aymeric Histace[3], Einas Abou Ali[1], Olivia Pietri[1], Aymeric Becq[1,2], Guy Houist[1], Isabelle Nion-Larmurier[1], Marine Camus[1,2], Christian Florent[1,2], Xavier Dray[1,2,3]

Institutions
1   Assistance Publique-Hôpitaux de Paris (APHP), Department of Hepatology and Gastroenterology, Saint Antoine Hospital, Paris, France
2   Sorbonne University, Paris, France
3   ETIS, Université de Cergy-Pontoise, ENSEA, CNRS, 95014 Cergy-Pontoise Cedex, France

Corresponding author
Xavier Dray, MD, PhD, Endoscopy Unit, Sorbonne University & APHP Saint Antoine Hospital, 184 rue du Faubourg Saint Antoine, 75012 Paris, France
Fax: +0033-149-282970
xavier.dray@aphp.fr

ABSTRACT

Background and study aims Capsule endoscopy (CE) is the preferred method for small bowel (SB) exploration. Its diagnostic yield can be reduced by poor mucosal visualization. We aimed to evaluate three electronic parameters – colorimetry, abundance of bubbles, and brightness – to assess the adequacy of mucosal visualization of SB-CE images.

Patients and methods Six-hundred still images were randomly extracted from 30 complete and normal SB-CEs. Three experts independently evaluated these images according to a 10-point assessment grid. Any frame with a mean score above seven was considered adequately cleansed. Each image was analyzed electronically according to the three preset parameters, individually and then combined, with the experts' score as reference. A random forests methodology was used for machine learning and testing.

Results The combination of the three electronic parameters achieved better discrimination of adequately from inadequately cleansed frames as compared to each individual parameter taken separately (sensitivity 90.0 % [95 %C.I. 84.1 – 95.9], specificity 87.7 % [95 %C.I. 81.3 – 94.2]).

Conclusion This multi-criterion score constitutes a comprehensive, reproducible, reliable, automated and rapid cleansing score for SB-CE frames. A patent is pending at the European patent office.

Introduction

Capsule endoscopy (CE) is the preferred method for small bowel (SB) exploration. Despite technological improvements, diagnostic yield of CE can be reduced by poor mucosal visualization secondary to the presence of residue, bile, bubbles in the digestive lumen, or as a result of insufficient or excessive brightness. The CE device has neither washing nor suctioning capabilities. Given these limitations, the quality of bowel preparation is of the utmost importance. There is no consensus regarding the preparation regimen for CE, mainly because there is no validated tool to assess the quality of mucosal visualization, contrary to colonoscopy for which the Boston Bowel Preparation Scale

(BPPS) is widely used [1]. Multiple bowel preparation regimens have been tested, aiming to improve mucosal visualization, with conflicting results [2].

Qualitative and quantitative clinical scores aiming at the assessment of mucosal visualization quality in CE have been established, with poor reproducibility, mostly due to the fact that they are based on evaluation of thousands of CE still frames [3 – 8]. Brotz et al. proposed a 10-point grading scale comprising the following five items: percentage of mucosa visualized, presence of fluid and debris, presence of bubbles, bile and chime staining, and brightness [4]. This score is not validated but is widely used in the field of research. Computed algorithms allowing automatic assessment of cleanliness of the SB

during CE have been developed. These algorithms are rapid and perfectly reproducible. Van Weyenberg et al. made a proof of concept with a score based on the ratio of color intensities of the red over green (R/G) channel of the tissue color bar of CE video segments [9]. The concept of R/G ratio is based on the fact that a properly visible mucosa is associated with red colors whereas a fecal-contaminated digestive lumen is associated with green. This approach was also used by Abou Ali et al. at the still frame level [10]. A R/G pixel ratio over 1.6 was found to yield a sensitivity and specificity of 91% defining an adequate SB visualization. Pietri et al. recently developed a computer algorithm based on a grey-level of co-occurrence matrix (GLCM) detector strategy, also yielding high diagnostic performances (sensitivity and specificity of 95%) in terms of assessing the abundance of bubbles in SB-CE still frames [11].

The main objective of this study was to evaluate the diagnostic performance of three computerized parameters (R/G ratio, bubble abundance, and brightness), individually and combined, to assess the quality of SB mucosa visualization in third-generation CE still frames.

Patients and methods

The study methodology is shown in a flowchart in ▶ Fig. 1.

Still frame selection

Patients who had undergone third-generation SB-CE (Pillcam SB3, Medtronic, Minnesota) in the setting of obscure gastrointestinal bleeding (OGIB) at Saint-Antoine Hospital, Paris, were included in the study. All procedures were complete and normal, meaning that the entire SB was examined and no lesion was observed. Incomplete and abnormal SB-CE were not considered for inclusion. SB-CE videos were deidentified. According to the French Data Protection Authority, no institutional ethical committee approval was required for this retrospective analysis of de-identified SB frames. Finally, 600 SB still frames were randomly extracted from these video sequences.

Expert analysis

Three SB-CE expert readers (with more than 1000 CE readings) analyzed the 600 extracted still frames independently, using a grid based on the quantitative scale by Brotz et al. [4]. Two points were allotted to each of the five items (mucosa visibility, brightness, bubbles, bile and chime, liquids and residues abundance) (▶ Table 1) with an overall score per still frame ranging from 0 to 10. A still frame was categorized as being of good visual quality when the mean of the three experts' scores was
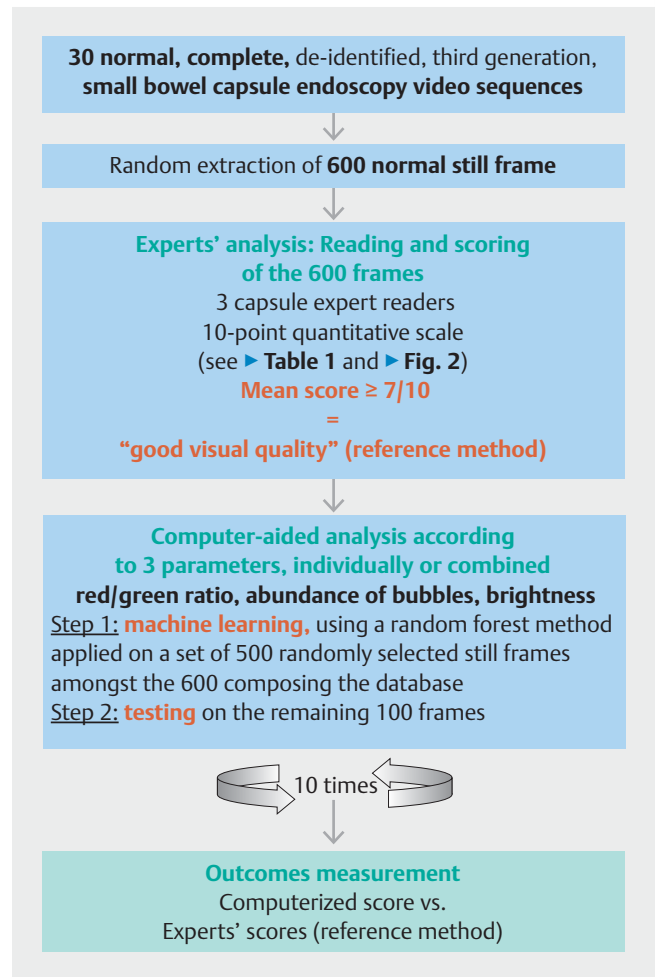

▶ **Fig. 1** Study methodology flowchart.

≥7/10. This threshold is based on the distribution curve of the quantitative score developed by Botz et al. [4]. By projection, regarding each item of the grading scale taking individually, a mean score ≥1.4/2 was considered to be of good visual quality as well.

Computer-aided and statistical analysis

Computerized analyses of the same 600 still frames were conducted using the MATLAB software (MathWorks, Natick, Massachusetts, United States). For each still frame, the following parameters were analyzed: (a) R/G ratio; (b) abundance of

▶ **Table 1** Quality of mucosal visualization grading scale, based on Brotz et al. quantitative scale [4].

Points	% of mucosa visualized	Liquids and residues abundance	Bubbles abundance	Chime/bile abundance	Reduction of brightness
0	<80%	Significant	Significant	Significant	Significant
1	80–89%	Moderate	Moderate	Moderate	Moderate
2	≥90%	Minimal/mild	Minimal/mild	Minimal/mild	Minimal/mild
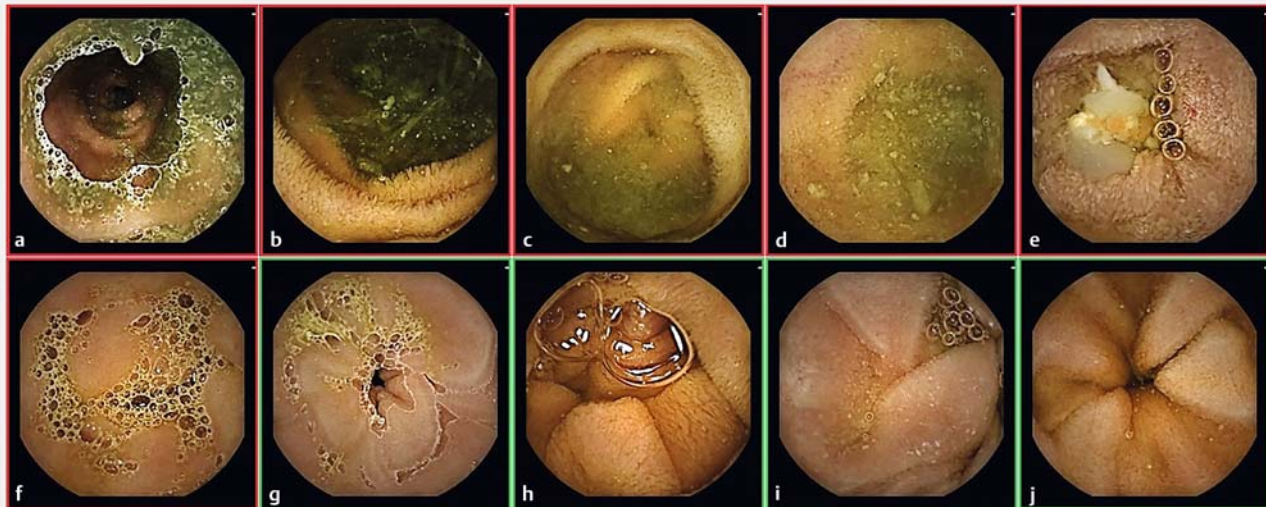
► **Fig. 2** Cleanliness of various small bowel still frames based on the 10-point grading scale (► Table 1) adapted from Brotz et al. [4] quantitative scale. Red frames: inadequate mucosal visualization (score <7/10) images with score of **a** 1, **b** 2 **c** 3, **d** 4, **e** 5, **f** 6. Green frames: adequate mucosal visualization (score ≥7/10) images with score of **g** 7, **h** 8, **i** 9, **j** 10.

► **Table 2** Experts' analysis of small bowel capsule endoscopy still frames, using the 10-point grading scale (► Table 1) based on Brotz et al. quantitative scale [4].

	Inadequate mucosal visualization (total score <7)	Adequate mucosal visualization (total score ≥7)
Expert 1 (n of images/%)	357 (59.5%)	243 (40.5%)
Expert 2 (n of images/%)	380 (63%)	220 (37%)
Expert 3 (n of images/%)	364 (61%)	236 (39%)
Mean score of 3 experts (n of images/%)	379 (63%)	221 (37%)

bubbles based on a GLCM detector strategy; and (c) brightness index.

GLCM (Grey-Level Co-occurrence Matrix) is an algorithm that permits definition of an image over a matrix showing the co-occurrence of gray-level by pair of pixels in a given neighborhood. From this matrix, the contrast parameter is computed as proposed by Haralick [12]. Index brightness concept is based on the fact that a variation of luminance or color allows an image to be discernable.

We used the "random forests method" based on decision trees combination [13]. In the statistical learning area, decision trees describe how to classify a set of complex data (such as a population, in our case, series of images) according to a set of discriminating variables (here, R/G ratio, GLCM contrast, brightness) and according to an objective set in terms of class number (here, images of "good" or "bad" quality of visualization) based on a ground truth (here, the mean scores of experts). To ensure better performance stability of this type of algorithm, decision tree forests perform learning on multiple decision trees driven on slightly different subsets of data created on the basis of the 500 images considered (strategy known as "bootstrapping" [14]).

The first step consisted of an automated learning process, applied on a set of 500 randomly selected still frames among the 600 composing the database. The second step consisted of a computed measurement of the remaining 100 still frames. To ensure the statistical validity of these measures, both phases were repeated 10 times. Finally, computerized measures were compared to the experts' evaluation. A mean score of 7/10 was used as a threshold for adequate mucosal visualization. Performance of the computerized analysis, sensitivity (Se), specificity (Sp), and positive (PPV) and negative (NPV) predictive values were calculated for each of the three parameters individually and then combined, using the expert analysis as reference. Se and Sp were computed from ROC (receiver operating characteristic) curves obtained using the output of the algorithm for each of the 100 test images: a probability of being of "good" or "bad" visualization. More precisely for this study, we focused on the Se and Sp corresponding to the particular "operating point" of the ROC curve for which the optimal trade-off was obtained.

The inter-observer correlation was calculated using κ statistics.

▶ **Table 3** Interobserver reproducibility of the experts' analysis of small bowel capsule endoscopy still frames.

	Kappa coefficient
Expert 1-Expert 2	0.83
Expert 1-Expert 3	0.81
Expert 2-Expert 3	0.87

Study outcomes

The primary outcome of the study was the Se of the combined computerized score. The secondary outcomes were the Sp, PPV and NPV of the combined score, the diagnostic performances of each of the three parameters, and the time of analysis.

Results

Still frame dataset

Thirty patients with OGIB were selected. They all had a normal and complete SB-CE. Six-hundred still frames were randomly extracted from these 30 videos. Examples of selected still images are shown in ▶ **Fig. 2**.

Expert analysis

Among the still frames, 221 (37 %) were categorized as having adequate mucosal visualization (mean score ≥ 7) (▶ **Table 2**).

The interobserver correlation between the three experts ranged from good to excellent (κ coefficient ranging between 0.81 and 0.87) (▶ **Table 3**).

Computerized analysis and outcomes

The combination of the three parameters achieved the highest diagnostic performance, with better discrimination between adequately and inadequately cleansed still frames as compared to using two parameters combined or each parameter individually (▶ **Table 4** and ▶ **Fig. 3**). Computerized analysis combining all 3 parameters demonstrated a Se of 90.0 % (95 %CI [84.1 – 95.9]), a Sp of 87.7 % (95 %CI [81.3 – 94.2]), a PPV of 81.1 % (95
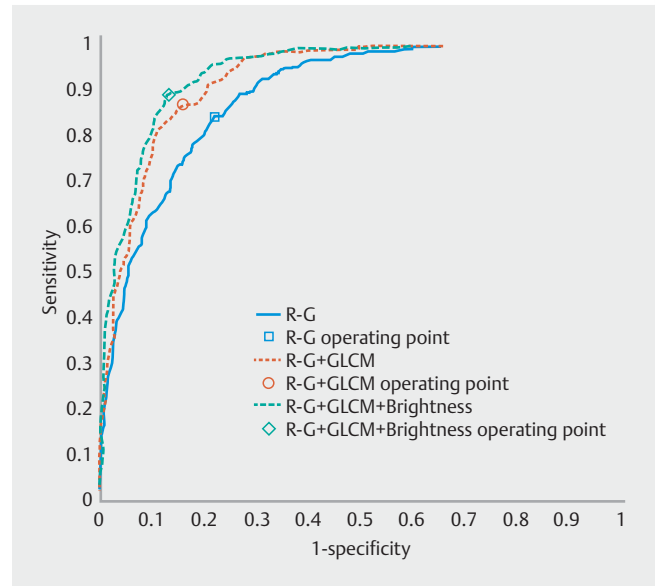
▶ **Fig. 3** Receiver operating characteristic curves for computerized parameters evaluating the cleanliness of small bowel capsule endoscopy still frames. R/G, red over green pixel ratio; GLCM, grey-level of co-occurrence matrix

%CI [73.3 – 88.7]), and a NPV of 93.7 % (95 %CI [88.9 – 98.4]). Reproducibility was optimal (κ coefficient = 1.0). Mean time required to analyze a still frame using the computerized three-parameter method was 34 ± 2 milliseconds using the MATLAB software. Extrapolated to a full-length CE video comprising 50,000 images, analysis time would take 28 minutes.

Discussion

We propose a multi-criterion computer-assisted algorithm to determine whether mucosal visualization is adequate on SB-CE still frames. The combination of the R/G ratio, abundance of bubbles, and brightness achieved a Se of 90.0 % and a Sp of 87.7 %, with optimal reproducibility, compared to human expert analysis. Further, the time of analysis was short.

▶ **Table 4** Diagnostic performances of computerized analysis to discriminate adequate from inadequate still frames.

Computerized parameter(s)	Sensitivity % (95 %CI)	Specificity % (95 %CI)
R/G ratio	84.1 (76.9; 91.2)	78.6 (70.6; 86.7)
Abundance of bubbles	79.6 (71.7; 87.5)	73.6 (64.9; 82.2)
Brightness	73.9 (65.4; 82.6)	78.4 (70.3; 86.4)
R/G ratio + abundance of bubbles	85.2 (78.3; 92.2)	86.3 (79.6; 93.1)
Abundance of bubbles + brightness	85.2 (78.2; 92.2)	79.0 (71.0; 87.0)
R/G ratio + brightness	86.1 (76.3; 92.9)	86.2 (79.4; 92.9)
R/G ratio + abundance of bubbles + brightness	90.0 (84.1; 95.9)	87.7 (81.3; 94.2)

R/G, red on green pixel ratio; CI, confidence interval

One of the strengths of this study was the solid analysis of still frames, performed using the 10-point grading scale, to obtain an adequate ground truth to which the computerized analysis was later compared. Three experienced capsule readers, blinded to the results of the computer-assisted analysis, performed this evaluation. A five-item standardized and precise scale allowed reliable clinical assessment of still frame quality of SB mucosa visualization [4]. Two of the three parameters used during computerized analysis had previously been evaluated and validated [10, 11]. We noted that the diagnostic performances of each individual parameter was lower than in previous studies [10, 11]. This difference can be explained by the fact that the expert analysis was more accurate, containing various parameters evaluated.

Some limitations of this study should be mentioned. First, we selected and evaluated still frames rather than video sequences. Most previously published studies aiming to build a SB-CE cleansing score were based on video sequence analysis by human readers. However, this could lead to heterogeneous results and variable conclusions [4, 6 – 8]. Our hypothesis was that a computer-assisted analysis would be objective, rapid, comprehensive, and reproducible. Taking into account that a computer-assisted scoring system would be based on a frame-by-frame analysis, we used still frames to build our ground truth. We used a grid based on quantitative scale developed by Brotz et al. [4] to evaluate still frames, although it was developed to evaluate video sequences. Interobserver correlation was good with κ coefficient >0.80. Additionally, we have started an evaluation of the multi-criterion score at the video level. Second, only normal SB-CEs performed in the setting of OGIB were included, which is not representative of the general population. We believe that any supposedly normal SB-CE should have adequate bowel preparation to be considered reliable. Thus, assessment of the quality of bowel preparation is much more significant in this setting, as opposed to when active bleeding or an abnormality is identified, regardless of the quality of preparation. Third, only cases of OGIB were selected, given it is the most prevalent indication for SB-CE, and because this group thus provides a homogenous population. Finally, we used the random forests method with two sets of 500 and 100 still frames. Despite an imbalance of datasets, this method allows a more solid automatic learning step and more reproducible results of the validation step.

Conclusion

In conclusion, this multi-criterion score constitutes a comprehensive, objective, reproducible, reliable, automated, and rapid test for evaluation of the level of cleanliness of SB-CE still frames. This automated score circumvents the subjectivity of qualitative or quantitative grading systems based on human reading. Further research is warranted to determine which proportion of adequately cleansed frames defines an acceptable quality of preparation of SB-CE in clinical practice, which would then be incorporated into CE software. For that purpose, a patent is pending at the European patent office.

Acknowledgement

Competing interests

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] Gkolfakis P, Tziatzios G, Dimitriadis GD et al. Meta-analysis of randomized controlled trials challenging the usefulness of purgative preparation before small-bowel video capsule endoscopy. Endoscopy 2018; 50: 671 – 683.3

[3] Ponte A, Pinho R, Rodrigues A et al. Review of small-bowel cleansing scales in capsule endoscopy: A panoply of choices. World J Gastrointest Endosc 20168: 600 – 609

[4] 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

[5] Albert J, Göbel C-M, Lesske J et al. Simethicone for small bowel preparation for capsule endoscopy: a systematic, single-blinded, controlled study. Gastrointest Endosc 2004; 59: 487 – 491

[6] Ninomiya K, Yao K, Matsui T et al. Effectiveness of magnesium citrate as preparation for capsule endoscopy: a randomized, prospective, open-label, inter-group trial. Digestion 2012; 86: 27 – 33

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

[8] 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

[9] 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

[10] Abou Ali E, Histace A, Camus M et al. Development and validation of a computed assessment of cleansing score for evaluation of the quality of small-bowel visualization in capsule endoscopy. Endosc Int Open 2018; 6: E646 – E651

[11] Pietri O, Rezgui G, Histace A et al. Development and validation of a highly sensitive and specific automated algorithm to evaluate the bubbles abundance in small bowel capsule endoscopy. Endosc Int Open 2018; 06: E462 – E469

[12] Haralick R, Shanmugam K, Dinstein I. Textural features for image classification. IEEE Transactions on Systems Man and Cybernetics 1973; 3: 610 – 621

[13] Breiman L. Random forests. Mach Learn 2001; 45: 5 – 32

[14] Kohavi R. A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence. Volume 2. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc; 1995: 1137 – 1143