Prospective Evaluation of Third-Generation Small Bowel Capsule Endoscopy Videos by Independent Readers Demonstrates Poor Reproducibility of Cleanliness Classifications

Xavier Dray, Guy Houist, Jean-Philippe Le Mouel, Jean-Christophe Saurin, Geoffroy Vanbiervliet, Chloé Leandri, Gabriel Rahmi, Clotilde Duburque, Julien Kirchgesner, Romain Leenhardt, et al.

Version of Record: https://www.sciencedirect.com/science/article/pii/S2210740120303788
Manuscript_3d1a8e3444d5ac3c4ee85a4f257f85e4

Prospective Evaluation of Third-generation Small Bowel Capsule Endoscopy Videos by Independent Readers Demonstrates Poor Reproducibility of Cleanliness Classifications

Xavier Dray (1), Guy Houist (2), Jean-Philippe Le Mouel (3), Jean-Christophe Saurin (4), Geoffroy Vanbiervliet (5), Chloé Leandri (6), Gabriel Rahmi (7), Clotilde Duburque (8), Julien Kirchgesner (9), Dr Romain Leenhardt (1), Franck Cholet (10)

(1) Sorbonne University, Center for Digestive Endoscopy, Hôpital Saint-Antoine, APHP, 184 rue du Faubourg Saint-Antoine, 75012 Paris, France

(2) Department of Hepato-Gastroenterology, Centre Hospitalier du Sud Francilien, 40 Avenue Serge Dassault, 91100 Corbeil-Essonnes

(3) Department of Hepato-Gastroenterology, Amiens University Hospital, Amiens, France

(4) Department of Hepato-gastroenterology, E. Herriot Hospital, Hospices civils de Lyon.

(5) Department of Hepato-Gastroenterology, L'Archet 2 hospital, Nice University hospital, Nice, F-06202 Cedex 3, France

(6) Department of Hepato-Gastroenterology, Cochin hospital, AP-HP, 75014, Paris

(7) Department of Hepato-Gastroenterology, Georges Pompidou European hospital, AP-HP, Paris Descartes University.

(8) Delegations for Clinical Research and Innovation - Department of Gastroenterology Lille Catholic hospitals - Lille Catholic University, Lille, France

(9) Sorbonne University and Department of Gastroenterology, Hôpital Saint-Antoine, Paris, France

(10) Department of Hepato-Gastroenterology, Brest University hospital, Bd Tanguy Prigent, 29200 BREST

Keywords: Capsule endoscopy; Gastrointestinal bleeding; Preparation; Small Bowel; Scoring.

Conflicts of interest

Pr Xavier Dray and Dr Romain Leenhardt are co-founders and shareholders of Augmented Endoscopy

Acknowledgement

This study has been made possible with the support of Norgine Pharma, France

ABSTRACT

Objective

The detection of lesions during small bowel (SB) capsule endoscopy (CE) depends on the cleanliness of the intestine. Quality reporting and comparison of different preparation methods require reliable scores. Three scores known as quantitative index (QI), qualitative evaluation (QE), and overall adequacy assessment (OAA), have been proposed to assess SB cleanliness, and are sometimes used in clinical practice and in clinical trials. However, none of these scores has received any external validation. The aim of our study was to re-assess the reproducibility of these three specific scores.

Methods

One-hundred-and-fifty-five complete third-generation SB-CE video recordings were extracted from a multicenter randomized controlled trial (PREPINTEST) which evaluated three modalities of SB preparation for CE. Three experts independently read the 155 SB-CE video recordings twice, in a random order, over 48-hour periods at 6-week intervals, using the QI, QE and OAA scores. Cohen's linearly weighted kappa coefficients were calculated to assess intra-observer and inter-observer agreements.

Results

Intra-observer reproducibility was fair to moderate, with kappa coefficients between 0.37 and 0.46 for QI, 0.41 and 0.51 for QE, 0.41 and 0.50 for OAA. Inter-observer reproducibility was fair to substantial correlations according to kappa coefficients between experts varying from 0.40 to 0.64, 0.29 to 0.65, and 0.52 to 0.71, for QI, QE and OAA, respectively.

Conclusions

QI, QE and OAA scores, currently used for evaluation of the quality of the preparation of SB-CE, are not sufficiently reproducible. Other scores or methods are therefore needed for SB-CE cleanliness assessment.

BACKGROUND

Small bowel (SB) capsule endoscopy (CE) is currently the reference examination for exploring the small bowel, with a diagnostic yield of 60% for obscure gastrointestinal bleeding (OGIB) (1) and 50% for suspected Crohn's disease (2). However, the ability to detect of lesions depends on how well the SB is prepared. Poor mucosal visualization, because of the presence of debris, bubbles, bile, fluids or insufficient light, may require the examination to be repeated (3). There is currently no consensus on the regimen and purge that must be carried out before SB-CE (4). The main difficulty in comparing the different preparations is the lack of a reliable and reproducible tool that can assess the cleanliness quality of the SB. Several cleanliness scales have been developed to try to solve this issue but they are poorly reproducible and time-consuming (5). In this vein, Brotz et al. described three SB-CE cleanliness assessment scores (6); these scores have often been used for clinical studies, first because the authors claim that they are validated (6), and also because experts have acknowledged that these scores were easy to use and were the most advanced to date in terms of performances for a global assessment of SB cleansing (5,7). However, these scores were slightly to substantially (but far from perfectly) reproducible and no external validation has ever been conducted. Hence, the aim of our study was to perform an external evaluation of these three scores.

PATIENTS AND METHODS

Selection of patients and videos

After Ethics Committee approval was obtained for such an ancillary study, SB-CE videos were selected from the Prepintest study (NCT01267981), a recent

multicenter, randomised controlled trial (RCT) assessing different modalities of SB cleansing regimens for CE (8). The study was approved by the ethical review board (Comité de Protection des Personnes Ouest 6 / 2010). Written, informed consent was obtained from each patient included in the study. The study protocol conforms to the ethical guidelines of the 1975 Declaration of Helsinki as reflected in a priori approval by the institution's human research committee. Eight-hundred-and-thirty-four patients with OGIB were included in the Prepintest trial. The patients were randomised in three groups (blocked randomization with randomly selected block sizes, stratified by center) : Prepa-1 (n = 277) standard diet; Prepa-2 (n = 284) standard diet + 500 mL PEG purge 30 minutes after SB-CE intake; Prepa-3 (n = 273) standard diet + 2000 mL PEG the night before + 500 mL PEG 30 minutes after SB-CE intake.(9)

For the current study, the inclusion criteria were the following: the patient was included in the Prepintest trial, there was a third-generation (SB3) CE recording, and there was a complete SB evaluation (with visualization of the colon). Patients were excluded if SB-CE was performed in another centre, if a first or second-generation (SB1 or SB2) CE was used, if a SB-CE recording was not available, or if the SB evaluation was incomplete (no colon frame recorded).

Cleanliness assessment of the small bowel

Three SB-CE expert readers (JCS, XD, JPLM) with 5 years or more of experience in SB-CE reading, with a total number of more than 500 CE readings and with commitments in SB-CE teaching and research, were selected. The included SB-CE recordings were edited at x32 speed using the Windows Media Player 10 software

(Microsoft, Redmond WA, USA). The experts independently read the included SB-CE recordings twice in a random order, over 48-hour periods at 6-week intervals, using the three grading systems for SB cleanliness assessment, as described in the study by Brotz et al. (6), as follows (**table 1**):

- A quantitative index (QI) based on the sum of five items on a 3-point scale (0, 1, 2) with a total score ranging from 0 to 10. The five items were the following: percentage of mucosa visualised, fluid and debris, bubbles, bile/chyme staining, and brightness.
- A first qualitative evaluation (QE) based on the QI items but simplified, grading the cleansing of the SB as poor, fair, good or excellent.
- A second qualitative index named the overall adequacy assessment (OAA), which is defined by the adequacy or inadequacy of the SB cleansing.

In each session, the first 20 video readings served as a training but their results were not recorded. However, these 20 sequences were re-read for data collection at the end of each session.

Statistical analysis

The quantitative variables were reported in mean and standard deviation (SD) values and the qualitative variables were reported in percentage values. Weighted Cohen's Kappa coefficients were calculated to assess intraobserver and interobserver agreements. Linear weighting was used to calculate the coefficient of the weighted κ-values (10). The interpretation of the weighted κ-values was performed using the table proposed by Landis and Koch (**Table 2**) (11).

RESULTS

Six-hundred-and-thirty-seven patients were selected from ~~the five most active centres of~~ the Prepintest trial. One-hundred-and-twenty-nine patients were excluded because of incomplete SB recording (n = 51) or because the videos could not be retrieved (n = 78). Among the remaining 516 remaining recordings, 361 second-generation CEs were excluded. Eventually, 155 patients with complete SB3-CE recordings were included (**Table 3**, patients' characteristics). Mean age was 61.43 ± 14.36 years. Seventy-two patients were male (46.45 %).

The 155 patients included in this ancillary study belonged to the following randomisation groups for SB preparation from the initial Prepintest study: group 1 (n=49; 31.61 %), group 2 (n=54; 34.84 %), and group 3 (n=52; 33.55 %). Nine patients out of 155 (5.81 %) received erythromycin. The mean SB transit time was of 199.73 ± 99.79 minutes. Sixty-seven patients out of 155 (43.22%) had at least one P1 or P2 SB lesion (with no statistical difference between the preparation groups). No adverse events were reported.

Intraobserver correlation

The intraobserver correlation was fair to moderate for the three scores, with operating points of weighted κ-values of ranging from 0.37 to 0.46 for QI, from 0.41 to 0.51 for QE, and from 0.41 to 0.50 for OAA **(Table 4)**. All QI items showed fair to moderate (0.21 to 0.60) correlations as well **(Table 5)**. Noticeably, the 'bile and chyme staining' subscore showed fair intraobserver reproducibility, with weighted κ-values varying from 0.21 to 0.29.

Interobserver correlation

The interobserver correlation was variable, from fair to substantial. Indeed, the mean [range] weighted κ-values were 0.52 [0.40-0.64], 0.48 [0.29-0.65], and 0.63 [0.52-

0.71], for QI, QE and OAA, respectively **(Table 6)**. Similarly, the correlation coefficients for the five QI items were highly variable (with weighted κ-values varying from 0.13 to 0.67) **(Table 7)**. Again, the 'bile and chyme staining' subscore was the least reproducible criteria, with weighted κ-values ranging between 0.13 and 0.34.

DISCUSSION

In this external validation study, intra- and interobserver correlations of the three cleanliness scores proposed by Brotz et al. for SB-CE (6) were found to be fair to moderate in most cases (weighted κ-values below 0.60) and occasionally substantial (between 0.61 and 0.80).

The current external validation study has a similar design compared with the initial study by Brotz et al. when it comes to assessing the proposed QI, QE, and OAA (6). Only slight methodological differences are to be acknowledged between the two studies, and to be recognized as strengths or limitations of our study compared to that of Brotz et al.(6) We used 155 SB3 video recordings (vs. 40 SB2 in the study by Brotz el al.). These 155 videos were read by three well-selected experts in our study, which is a limitation compared to the five readers in the study by Brotz et al. Patients received variable preparation regimens, sometimes with PEG (vs. standard diet only in the study by Brotz el al.). SB-CE recording readings were performed on single-view compressed accelerated (x32) videos (vs. four-view, native, 30 frames per second in the study by Brotz el al.), which is another potential limitation of our study. Both readings were doubled in a random order, at 6-week intervals (vs. 4-week intervals in the study by Brotz el al.). The entire SB cleanliness was scored (vs. the entire and specifically the distal SB in the study by Brotz el al.). The statistics used

weighted κ-coefficients (vs. unweighted in the study by Brotz el al.). We describe the range intervals as well (rather than the 95% confidence intervals [95%C.I.] based on five to ten measurements, which seems inappropriate, in the study by Brotz et al.)

The results were quite similar between the two studies. For comparison, for the entire SB, QI intraobserver correlations ranged between 0.37 and 0.46 (vs. [95%C.I.] 0.35 to 0.77, unweighted, in the study by Brotz el al.), and the interobserver correlations ranged between 0.40 and 0.64 (vs. [95%C.I] 0.22 to 0.67, unweighted); QE intraobserver correlations ranged between 0.41 and 0.51 (vs. [95%C.I] 0.18 to 0.55, unweighted), and the interobserver correlations ranged between 0.29 and 0.65 (vs. [95%C.I] 0.13 to 0.26, unweighted); OAA intraobserver correlations ranged between 0.41 and 0.50 (vs. [95%C.I] 0.29 to 0.83, unweighted); and the inter-observer correlations ranged between 0.52 and 0.71 (vs. [95%C.I] 0.31 to 0.50, unweighted).

Overall, the methods and results of the current external study are similar to those by Brotz et al. However, we disagree regarding the interpretation of these results by Brotz et al.

Two recent experts' reviews quoted the three scores proposed by Brotz el al. as validated for assessing the quality of cleanliness of the small intestine (5,7). We agree on the fact that these three scores are easy to use. However, based on poor reproducibility, for all three grading systems and for both intra- and interobserver correlations, we cannot conclude that any of these scores can be validated for clinical care or for research.

Our conclusion, there are still that unmet needs remain regarding the assessment of cleanliness of the SB in CE. There is room for research and improvement in this field. Both the study by Brotz et al. and the current study indicate that a human interpretation a full-length SB-CE recording (that encompasses a mean number of 50,000 frames) is hardly reproducible (and definitely tedious), even with a QI score that includes meaningful items (percentage of mucosa visualised, fluid and debris, bubbles, bile/chyme staining, and brightness). Artificial intelligence (computed-assisted diagnosis) may solve this issue in the future (13).

Tables

Table 1: Quantitative index, qualitative evaluation, and overall adequacy assessment, according to Brotz et al. (6)

Quantitative index ***Elements*** Percentage of mucosa visualized* Fluid and debris Bubbles Bile/chyme staining Brightness ***Score per element***¥ 2 = Minimal/mild impairment 1 = Moderate impairment 0 = Severe impairment
Qualitative evaluation ***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.
Overall adequacy assessment ***Adequate*** ***Inadequate***

* Severe < 80% = 0, moderate 80-89% = 1, minimal/mild = 2

¥ Total score = 0-10 ; higher score = superior cleansing

Table 2: Kappa test interpretation according to Landis and Koch (11).

k	Reliability
0 - 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost perfect

Table 3: Patients' characteristics

Male gender	72 (46.4%)
Age (years, mean ± SD)	61.4 ± 14.4
Lowest haemoglobin rate (g/dL, mean ± SD)	8.5 ± 2.0
Diabetes	39 (25.2%)
Medications Non-steroid anti-inflammatory drugs Anticoagulation Aspirin Clopidrogrel	 7 (4.5%) 20 (12.9%) 34 (21.9%) 9 (5.8%)
Iron supplementation	115 (74.2%)
Blood transfusion	73 (47.1%)

Table 4: Intra-observer agreements (Cohen's linearly weighted kappa coefficients, operating points) of three expert readers assessing the small bowel cleanliness of capsule endoscopy examinations using the three scores by Brotz et al. (6)

	Quantitative index (QI)	**Qualitative evaluation (QE)**	**Overall Adequacy Assessment (OAA)**
Expert 1	0.37	0.41	0.49
Expert 2	0.44	0.46	0.41
Expert 3	0.46	0.51	0.50

Table 5: Intra-observer agreements (Cohen's linearly weighted kappa coefficients, operating points) of three expert readers assessing the small bowel cleanliness of capsule endoscopy examinations using the 5 items of the qualitative index by Brotz et al. (6)

	Mucosal visualisation	**Fluid and debris**	**Bubbles**	**Bile/chyme staining**	**Brightness**
Expert 1	0.56	0.35	0.35	0.21	0.22
Expert 2	0.48	0.55	0.52	0.26	0.46
Expert 3	0.49	0.60	0.55	0.29	0.52

Table 6: Inter-observer agreements (Cohen's linearly weighted kappa coefficients, operating points) of three expert readers assessing the small bowel cleanliness of capsule endoscopy examinations using the three scores by Brotz et al. (6)

	Quantitative index (QI)	**Qualitative evaluation (QE)**	**Overall Adequacy Assessment (OAA)**
First Reading (at week-0, random order)			
Expert 1 Vs 2	0.49	0.48	0.52
Expert 1 Vs 3	0.57	0.53	0.70
Expert 2 Vs 3	0.55	0.65	0.66
Second Reading (at week-6, random order)			
Expert 1 Vs 2	0.40	0.29	0.52
Expert 1 Vs 3	0.46	0.43	0.67
Expert 2 Vs 3	0.64	0.53	0.71

Table 7: Inter-observer agreements (Cohen's linearly weighted kappa coefficients, operating points) of three expert readers assessing the small bowel cleanliness of capsule endoscopy examinations using the 5 items of the qualitative index by Brotz et al. (6)

	Mucosal visualisation	**Fluid and debris**	**Bubbles**	**Bile/chyme staining**	**Brightness**
First Reading (at week-0, random order)					
Expert 1 Vs 2	0.55	0.61	0.48	0.18	0.49
Expert 1 Vs 3	0.63	0.63	0.64	0.13	0.52
Expert 2 Vs 3	0.67	0.65	0.65	0.26	0.53
Second Reading (at week-6, random order)					
Expert 1 Vs 2	0.30	0.37	0.30	0.23	0.31
Expert 1 Vs 3	0.41	0.40	0.33	0.16	0.28
Expert 2 Vs 3	0.48	0.49	0.61	0.34	0.32

References

(1) Pennazio M, Spada C, Eliakim R et al. Small-bowel capsule endoscopy and device-assisted enteroscopy for diagnosis and treatment of small-bowel disorders: European Society of Gastrointestinal Endoscopy (ESGE) Clinical Guideline. Endoscopy 2015; 47:352-376

(2) Triester SL, Leighton JA, Leontiadis GI et al. A meta-analysis of the yield of capsule endoscopy compared to other diagnostic modalities in patients with non-stricturing small bowel Crohn's disease. Am J Gastroenterol. 2006;101:954–964.

(3) Jones B, Fleischer D, Sharma V et al. Yield of Repeat Wireless Video Capsule Endoscopy in Patients with Obscure Gastrointestinal Bleeding. The American Journal of Gastroenterology. 2005;100:1058-1064.

(4) Niv Y. Efficiency of bowel preparation for capsule endoscopy examination: a meta-analysis. World J Gastroenterol. 2008;14:1313–7.

(5) 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 2016;8:600-9.

(6) 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

(7) Spada C, McNamara D, Despott EJ, et al. Performance measures for small-bowel endoscopy: a European Society of Gastrointestinal Endoscopy (ESGE) Quality Improvement Initiative). Endoscopy. 2019;51(6):574-598

(8) F. Cholet, G. Rahmi, M. Gaudric et al. Does polyethylene glycol cleansing purge improve video capsule endoscopy diagnostic yield in obscure gastrointestinal bleeding, Endoscopy 2018;50(04): S18

(9) D'Halluin PN, Delvaux M, Lapalus MG et al. Does the "Suspected Blood Indicator" improve the detection of bleeding lesions by capsule endoscopy? Gastrointest Endosc. 2005;61(2):243-9.

(10) Brenner H, Kliebsch U. Dependence of weighted kappa coefficients on the number of categories. Epidemiology 1996;7:199–202

(11) Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.

(12) 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;43:406-11.

(13) Leenhardt R, Souchaud M, Houist G, Le Mouel JP, Saurin JC, Cholet F, Rahmi G, Leandri C, Histace A, Dray X. A Neural Network-based Algorithm for Assessing the Cleanliness of Small Bowel during Capsule Endoscopy. Endoscopy 2020 (in press)