Thieme

Validating bowel preparation scales

OPEN ACCESS

Authors
Valérie Heron[1], Robin Parmar[1], Charles Ménard[2], Myriam Martel[1], Alan N. Barkun[1,3]

Institutions
1 Division of Gastroenterology, The McGill University Health Center, Montreal General Hospital, McGill University, Montréal, Québec, Canada
2 Department of Medicine, University of Sherbrooke, Sherbrooke, Québec, Canada.
3 Division of Clinical Epidemiology, The McGill University Health Center, Montreal General Hospital, McGill University, Montréal, Canada

submitted 20.2.2017
accepted after revision 31.7.2017

Bibliography
DOI https://doi.org/10.1055/s-0043-119749 |
Endoscopy International Open 2017; 05: E1179–E1188

ISSN 2364-3722

Corresponding author
Alan Barkun, MD, Montreal General Hospital, 1650 Cedar Avenue, #D7-148, Montreal, Quebec, H3G 1A4, Canada
Fax: +1-514-934-8375
alan.barkun@muhc.mcgill.ca

ABSTRACT

Background and study aim Few scales assessing bowel preparation quality have been validated, and direct between-scale comparisons remain scarce. The aim of the study was to compare inter- and intra-rater reliability, predictive abilities for clinical outcomes, and ease of use for each scale.

Methods Colonoscopy video recordings highlighting five colonic segments after washing were viewed independently by three physicians, and cleanliness was evaluated using the Boston Bowel Preparation Scale (BBPS), the Chicago Bowel Preparation Scale (CBPS), and the Harefield Cleansing Scale (HCS) in randomized order. Kappa or intraclass correlations quantified intra- and inter-rater reliability. Ease of use was evaluated (1 – 10 scale, 1 = easy), as were associations between scores, adenoma detection, and adequacy of preparation to exclude lesions ≥ 5 mm.

Results Among 83 colonoscopy videos, indications included screening or surveillance in 72.3 %. Mean (± SD) scores of the respective three raters were 5.17 ± 1.57, 6.49 ± 1.48, and 5.12 ± 1.21 for BBPS, and 23.73 ± 6.01, 28.39 ± 5.47, and 24.75 ± 5.83 for CBPS, while successful HCS scores (grade A or B) were given for 76 %, 89 %, and 63 % of examinations. Intra-rater reliability ranges were 0.88 – 1.00, 0.83 – 1.00, and 0.62 – 1.00 for BBPS, CBPS, and HCS, respectively. Similarly, inter-rater reliability ranges were 0.50 – 0.79, 0.64 – 0.83, and 0.28 – 0.52, respectively. Sources of disagreement included varying rater strictness, which was possibly most marked for preparations rated as intermediate. Overall, associations between preparation scores and adenoma detection lacked statistical significance.

Conclusion The BBPS and CBPS showed the best inter- and intra-rater reliability, and the BBPS was considered the easiest to use. Further studies are needed to determine an optimal adequacy threshold for these scales, with the goal of predicting clinical outcomes and determining the appropriate interval to the next colonoscopy.

Introduction

Colorectal cancer (CRC) is one of the most prevalent cancers in both men and women worldwide [1]. Routine screening using colonoscopy has contributed to significantly decreasing both the incidence of CRC and mortality from this disease [2 – 4]. The quality of the colonoscopy is critical for realizing this potential, with one of the most important factors ensuring a high-quality examination being the adequacy of the bowel preparation [5, 6]. The US Multi-Society Task Force has recommended early interval follow-up to the next colonoscopy if the preparation does not allow for the detection of polyps greater than 5 mm in size [7].

Over the years, several scales have been developed in order to better quantify for clinical or research purposes the adequacy of cleansing attributable to a bowel preparation [8 – 16]. However, few of these have been formally validated to guide clinical management. There exist very limited data to suggest an appropriate time interval to repeat colonoscopy based on quality of bowel preparation. Moreover, a recent meta-analysis

suggests that preparations of intermediate-level cleanliness may result in outcomes similar to those of cleaner preparations [17]. In addition, data on direct head-to-head comparisons of the performance and perceived simplicity of the different scales remain scarce [16].

We have therefore chosen to compare three of the most commonly used scales: the Boston Bowel Preparation Scale (BBPS) [10–13], the Chicago Bowel Preparation Scale (CBPS) [15], and the Harefield Cleansing Scale (HCS) [14]. For each of these scales, intra-rater and inter-rater reliability were assessed using a standardized set of colonoscopy video clips, as well as predictive abilities for clinical outcomes (adenoma detection and ability to detect lesions ≥ 5 mm). The ease of use of each scale was also evaluated.

Methods

Study population

Our study took place in an urban, university-affiliated hospital where over 6100 colonoscopies are performed each year by 15 endoscopists, nine of whom are gastroenterologists. The study included all patients above the age of 18 years who required a colonoscopy and who were able to provide informed consent. Patients were excluded if they had a previous segmental colectomy or if video recordings of five colonic segments could not be achieved.

Colonoscopy video recordings

Representative video clip recordings from 83 complete colonoscopies to the cecum were prospectively collected between 15 July and 6 August 2014 using our endoscopy unit's EndoWorks software (Olympus Corporation of the Americas, Center Valley, Pennsylvania, USA). For each colonoscopy, at least five video clips were obtained: at least one distinct video clip per colonic segment (right, transverse, left, sigmoid, and rectum). These were recorded after optimal washing had been carried out. Video clips were labeled by location (segment) for use by the raters. Video recordings intentionally included both very clean and very poor preparations to ensure that the breadth of possible cleansing outcomes was represented, though the majority of clips were of intermediate-level cleanliness. Each video was approved by two trained physicians based on quality of the images and representativeness of the colonoscopy.

Patient demographic and endoscopic data collection

Demographic information of patients such as date of birth, sex, indication for colonoscopy, and details of colon preparation was recorded at the time of colonoscopy, as was procedural information (i. e. withdrawal time, complications, endoscopic findings, polyp removal, and recommended interval to next colonoscopy). Corresponding pathology reports were also collected to determine the number of adenomas removed.

Bowel preparation scales

We selected three previously validated bowel preparation scales that do not specifically require a fluid score, in order to allow the inclusion of videos after optimal washing. The BBPS evaluates three colonic segments (right, transverse, left) on a scale of 0 to 3 with a total score out of 9, 9 being the cleanest [10–13]. The CBPS scores each of these same segments on a scale of 0 to 12 for a total score out of 36, with higher scores representing cleaner preparations [15]. However, for the purposes of our study, we rated each segment from 0 to 11 with a total CBPS score out of 33 in order to exclude the assessment of the need for washing (one point per segment), given that our videos were obtained after optimal washing had already been achieved. The HCS provides a score of 0 to 4, 4 being clean, for each of the five colonic segments (right, transverse, left, sigmoid, rectum). Based on the scores attributed to each segment, the overall preparation then receives a grade A, B, C or D. Grades A and B are interpreted as successful cleansing, whereas C and D are considered unsuccessful [14].

Reliability assessment

All colonoscopy video recordings were viewed by three physicians (one full-time staff gastroenterologist [C. M.] and two junior gastroenterology trainees with extensive exposure to colonoscopies and experience rating preparations [V. H., R. P.]). These three physicians independently evaluated the bowel preparation quality of each examination using the BBPS, CBPS, and HCS in randomized order. Five of these colonoscopies were selected for assessment of intra-rater reliability. For this purpose, the video clips of each colonic segment for these five colonoscopies were viewed and rated again by each rater 1 month after initial evaluation.

Each rater was provided with a detailed written description of the three scales in order to optimize agreement (description available upon request). The juniors, but not the senior rater, were also provided with video clips of each of the five colonic segments illustrating varying levels of bowel cleanliness. However, no formal calibration meeting was held. A calibration image was used by all three raters in order to standardize the appreciation of a 5 mm measurement in the colon. In addition, raters were instructed to provide a rating based solely on the contents of the video, assuming that optimal cleansing had been achieved and that the images were representative of the entire colonoscopy. For BBPS and CBPS, videos of the sigmoid and rectum were considered to be part of the left colon, and were rated accordingly. Bowel preparation scores were compiled using standardized data collection sheets.

Validity assessment

All raters were blinded to the clinical outcomes of colonoscopies included in the study. Based on the colonoscopy videos, raters commented as to whether or not the preparation was adequate to exclude lesions ≥ 5 mm. Construct validity was assessed using rater's opinion on adequacy to detect lesions ≥ 5 mm as well as adenoma detection rates. Although the US Multi-Society Task Force recommendation refers to the detec-

tion of lesions greater than 5 mm [7], the Canadian Association of Gastroenterology quality guidelines are not as clear as to the inclusion of a 5 mm lesion [18]. We thereby adopted a cutoff that included lesions 5 mm in size. Ease of use of each bowel preparation scale was also evaluated by each rater on a 1 – 10 point Likert scale, 1 being very simple to use and 10 being very complicated.

Exploratory analyses of heterogeneity in inter-rater reliability

A series of pre-planned exploratory analyses were carried out in an observational fashion in an attempt to understand possible reasons for heterogeneity in BBPS scoring. These included assessing reliability across high-, intermediate-, and low-scored preparations, and additional scoring criteria for determining the interval to the next colonoscopy [19]. The choice of the BBPS among other scales for this exploratory analysis was based on the results of a recently published systematic review that identified the BBPS as being the most validated bowel preparation scale in the literature [16]. Furthermore, there is some controversy in the literature concerning the optimal BBPS cutoff score to allow for routine surveillance interval follow-up.

Reliability was also assessed for the subgroups of preparations considered adequate to detect 5 mm lesions and those considered inadequate to do so according to the senior rater.

Statistical analysis

Both intra-rater and inter-rater reliability were quantified using kappa scores for nominal values with 95 % confidence intervals (CIs) following the Landis – Koch benchmarks [20]. The strength of agreement of the kappa values was characterized as follows: < 0 poor, 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. Interclass correlation coefficient (ICC) was calculated for continuous variables with a two-way random average measure and reported with 95 %CI [21]. The ICC coefficient was characterized as follows: values below 0.4 represent poor reliability, values above 0.75 represent excellent reliability, and values between 0.4 and 0.75 represent fair-to-good reliability [22].

A sample size of 83 subjects for inter-rater variability assessment had been previously calculated for a separate ongoing study (Barkun A. N., personal communication) assessing a new preparation evaluation scale based on expected ICC values of 0.70 for senior gastroenterologists and 0.80 for junior trainees both with an approximate expected 95 %CI width of ± 0.10. Adopting similar estimates, with three raters (two juniors and one senior) with the lowest expected ICC (0.70), the approximated sample size based on Giraudeau and Mary [23] is 63 subjects.

Associations between scores and adenoma detection as well as between scores and raters' opinions on adequacy of preparation to exclude lesions ≥ 5 mm were assessed for independent samples with a chi-squared test (or Fisher's exact test) for categorical variables and a *t* test for continuous variables. A 2-sided *P* value threshold of 0.05 was adopted for statistical significance. All statistical analyses were performed using SAS version 9.3 (SAS Institute Inc., Cary, North Carolina, USA).

Finally, exploratory receiver operating characteristic curves were used to evaluate BBPS and CBPS score thresholds correlating with the ability to detect lesions ≥ 5 mm in size based on the assessment by the senior rater. Sensitivity and specificity were calculated according to the cutoff. The HCS did not lend itself to this analysis as it was designed as a categorical rather than continuous scoring system.

Ethical considerations

Approval for this study was obtained from the ethics committee of the McGill University Health Centre, and written consent was obtained from each patient prior to the recording of colonoscopy videos.

Results

Patient demographic and endoscopic data

A total of 83 colonoscopies, each represented by at least five video clips, were independently reviewed by the three raters. Five of these colonoscopies were used for intra-rater variability. Average video clip duration was 23.9 ± 12.6 seconds per colonic segment. The patient population included 41 women (49.4 %), and mean age was 64.4 ± 12.4 years. Colonoscopy was performed for screening or surveillance purposes in 72.3 % of patients. Other indications included rectorrhagia above the age of 40 years (6.0 %), anemia (6.0 %), probable active inflammatory bowel disease (3.6 %), and a variety of other indications (data available upon request). The bowel preparations administered were PICO-SALAX (Ferring Pharmaceuticals Inc., North York, Ontario, Canada; 67.5 %) or GoLYTELY (Braintree Laboratories, Inc. Braintree, Massachusetts, USA; 32.5 %). Magnesium citrate was used as an adjuvant in 48 %. Mean withdrawal time recorded was 7 minutes and 38 seconds. A repeat colonoscopy was recommended by the endoscopist because of poor preparation in 4.8 % of cases. Adenomas were removed in 22.9 % of these highly selected patients.

Reliability testing

Mean (±SD) scores for each of the three raters were 5.17 ± 1.57, 6.49 ± 1.48, and 5.12 ± 1.21 using the BBPS. Mean scores for the CBPS were 23.73 ± 6.01, 28.39 ± 5.47, and 24.75 ± 5.83. For the HCS, a successful cleansing score (grade A or B) was given for 76 %, 89 %, and 63 % of examinations, respectively.

Intra-rater reliability ranged between 0.88 and 1.00 for the BBPS, 0.83 and 1.00 for the CBPS, and 0.62 and 1.00 for the HCS (▶ **Table 1**). Similarly, inter-rater reliability ranged between 0.50 and 0.79 for the BBPS, 0.64 and 0.83 for the CBPS, and 0.28 and 0.52 for the HCS (▶ **Table 2**). Overall inter-rater reliability for each of these scales was 0.75, 0.80, and 0.39, respectively.

Validity testing

Mean scores according to the pre-defined clinical outcomes (adenoma detection and ability to detect 5 mm lesions) are shown in ▶ **Table 3**.

▶ **Table 1** Intra-rater reliability.

	Rater 1	Rater 2	Rater 3
BBPS[1], ICC, % (95 %CI)	0.88 (– 0.54 to 0.99)	1.00	0.99 (0.89 to 1.00)
CBPS[1], ICC, % (95 %CI)	0.97 (0.76 to 1.00)	0.83 (– 0.59 to 0.98)	1.00 (0.98 to 1.00)
HCS[2], κ (95 %CI)	0.62 (0.00;1.00)	1.00	1.00

ICC, interclass correlation coefficient; CI, confidence interval; BBPS, Boston Bowel Preparation Scale; CBPS, Chicago Bowel Preparation Scale; HCS, Harefield Cleansing Scale.
[1] Total scores.
[2] Scores A or B.

▶ **Table 2** inter-rater reliability.

	Overall raters	Rater 1 vs. Rater 2	Rater 1 vs. Rater 3	Rater 2 vs. Rater 3
BBPS[1], ICC, % (95 %CI)	0.75 (0.43 to 0.87)	0.71 (– 0.13 to 0.90)	0.79 (0.68 to 0.87)	0.50 (– 0.10 to 0.75)
CBPS[1], ICC, % (95 %CI)	0.80 (0.59 to 0.89)	0.70 (– 0.01 to 0.88)	0.83 (0.73 to 0.89)	0.64 (0.25 to 0.81)
HCS[2], κ (95 %CI)	0.39 (0.26 to 0.51)	0.52 (0.31 to 0.74)	0.47 (0.28 to 0.67)	0.28 (0.10 to 0.46)

ICC, interclass correlation coefficient; CI, confidence interval; BBPS, Boston Bowel Preparation Scale; CBPS, Chicago Bowel Preparation Scale; HCS, Harefield Cleansing Scale.
[1] Total scores.
[2] Scores A or B.

▶ **Table 3** Scores according to clinical outcomes.

	Scores associated with adenoma detection			Scores associated with ability to detect 5 mm lesions		
	≥ 1 Adenoma detected (n = 19)	No adenoma detected (n = 62)	*P* value	Adequate to detect ≥ 5 mm	Inadequate to detect ≥ 5 mm	*P* value
BBPS[1], mean ± SD						
▪ Rater 1	5.2 ± 1.2	5.0 ± 1.7	0.25	6.3 ± 1.3	4.5 ± 1.3	< 0.01
▪ Rater 2	6.4 ± 1.6	6.8 ± 0.8	0.26	7.1 ± 1.1	5.1 ± 1.3	< 0.01
▪ Rater 3	5.4 ± 1.0	5.0 ± 1.3	0.29	6.1 ± 0.8	4.5 ± 1.0	< 0.01
CBPS[1], mean ± SD						
▪ Rater 1	25.3 ± 4.2	23.2 ± 6.5	0.11	28.5 ± 4.0	20.9 ± 5.2	< 0.01
▪ Rater 2	30.4 ± 2.2	27.7 ± 6.1	< 0.01	30.8 ± 1.9	23.2 ± 7.0	< 0.01
▪ Rater3	26.5 ± 4.6	24.3 ± 6.2	0.15	30.1 ± 3.0	21.2 ± 4.3	< 0.01
HCS[2], % (95 %CI)						
▪ Rater 1	79.0 (58.8 to 99.1)	75.5 (64.8 to 86.8)	0.99	90.3 (79.3 to 100.0)	67.3 (54.1 to 80.5)	0.02
▪ Rater 2	100.0	87.1 (78.5 to 95.7)	0.19	98.2 (94.6 to 100.0)	72.0 (53.1 to 90.9)	< 0.01
▪ Rater 3	73.7 (51.9 to 95.5)	60.0 (47.1 to 72.2)	0.27	93.9 (85.4 to 100.0)	42.0 (27.8 to 56.2)	< 0.01

ICC, interclass correlation coefficient; CI, confidence interval; BBPS, Boston Bowel Preparation Scale; CBPS, Chicago Bowel Preparation Scale; HCS, Harefield Cleansing Scale.
[1] Total scores.
[2] Scores A or B.

BBPS scores of colonoscopies that led to the detection of an adenoma ranged between 5.2 ± 1.2 and 6.4 ± 1.6; when no adenoma was detected, BBPS scores ranged between 5.0 ± 1.7 and 6.8 ± 0.8. CBPS scores associated with adenoma detection ranged between 25.3 ± 4.2 and 30.4 ± 2.2; CBPS scores for preparations of colonoscopies where adenomas were not found ranged from 23.2 ± 6.5 to 27.7 ± 6.1. Adequate Harefield scores (A or B) were given in 73.7 % to 100 % of preparations where the colonoscopy ultimately led to the detection of one or more ade-

▶ **Fig. 1** Ease of use of bowel preparation scales (1 = very simple to use; 10 = very complicated). BBPS, Boston Bowel Preparation Scale; CBPS, Chicago Bowel Preparation Scale; HCS, Harefield Cleansing Scale.

nomas; 60 % – 87.1 % of colonoscopies that did not lead to the detection of adenomas received an adequate score.

Preparations that were considered adequate to detect lesions ≥ 5 mm in size were given BBPS scores between 6.1 and 7.1 compared with scores of 4.5 to 5.1 in preparations considered inadequate to do so ($P < 0.01$). CBPS scores between 28.5 and 30.8 were attributed to preparations adequate to detect lesions ≥ 5 mm, whereas preparations considered inadequate to do so received scores between 20.9 and 23.2 ($P < 0.01$). Preparations judged adequate to detect lesions ≥ 5 mm in size received an adequate HCS score (A or B) in 90.3 % – 98.2 % of cases compared with 42.0 % – 72.0 % when the preparation was not considered adequate to detect lesions ≥ 5 mm ($P < 0.01$ to 0.02).

Associations between bowel preparation scores and withdrawal time, as well as the interval to repeat colonoscopies, were also assessed (data available in ▶ **Appendix 1**). No significant differences were observed in BPS scores based on withdrawal time of ≤ 7 minutes vs. > 7 minutes. For all three scales, tendencies toward higher scores were observed when the recommended interval to next colonoscopy was greater than 1 year, though these were not statistically significant except for CBPS raters 1 and 2. Receiver operating characteristic curves identified an optimal BBPS threshold score of 6 (area under the curve [AUC] 0.91; sensitivity 90.9 %; specificity 90.0 %) and CBPS threshold score of 26 (AUC 0.96; sensitivity 93.9 %; specificity 88.0 %) correlating with the ability to detect lesions ≥ 5 mm (▶ **Appendix 2**).

Ease of use scores ranged between 2 and 3 for BBPS, 3 and 7 for CBPS, and 5 and 7 for HCS (1 being most simple to use and 10 being most complicated; ▶ **Fig. 1**).

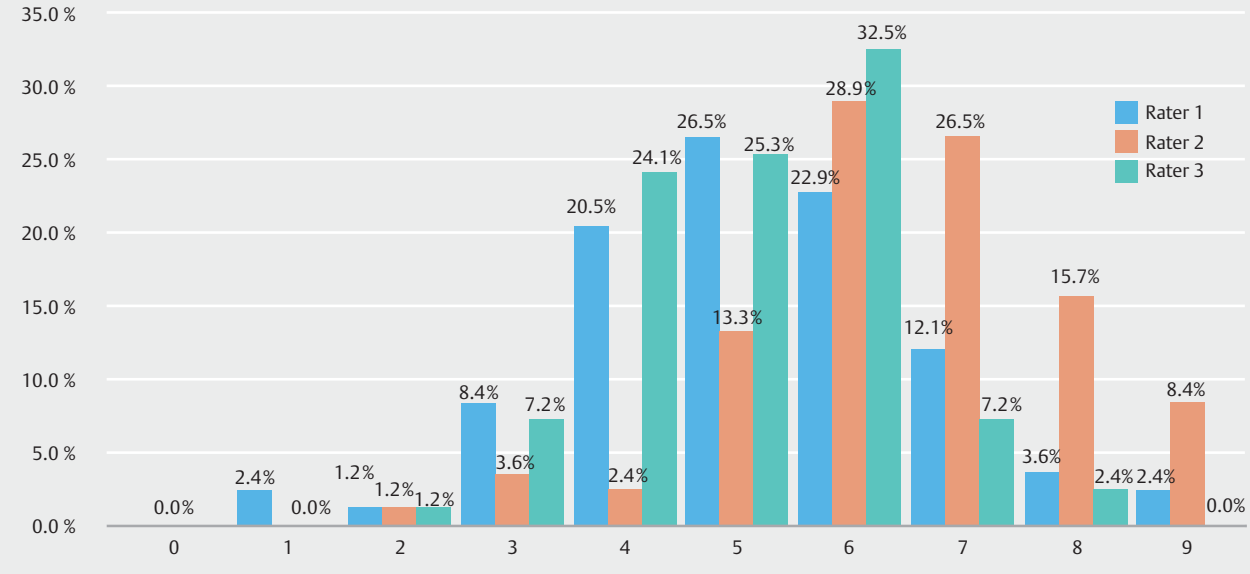
Exploratory analysis

The distribution of BBPS scores attributed by each rater was assessed and is illustrated in ▶ **Fig. 2**.

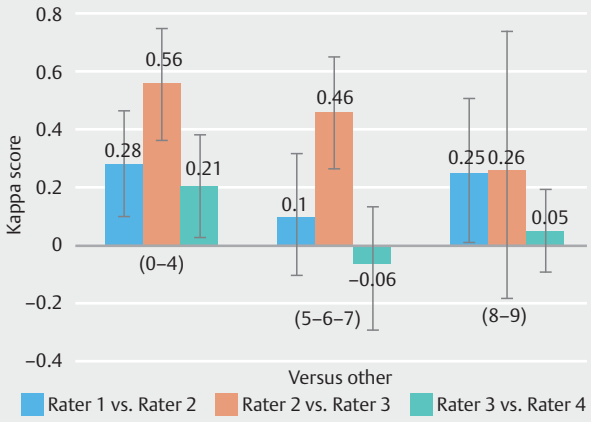
Inter-rater reliability was assessed for total scores < 6 vs. ≥ 6 as well as for total scores < 7 vs. ≥ 7. These thresholds were chosen based on previous reliability testing of the BBPS [10]. Inter-rater reliability for BBPS score < 6 vs. ≥ 6 ranged between 0.18 and 0.53. When using a cutoff of BBPS < 7 vs. ≥ 7, inter-rater reliability ranged between 0.09 and 0.35.

Based on a recent standardized definition of adequate BBPS for 10-year follow-up, inter-rater reliability was also assessed for preparations meeting the criteria of total BBPS ≥ 6 and all segment scores ≥ 2 [19]. With this set of criteria, inter-rater reliability ranged between 0.31 and 0.59.

We assessed inter-rater reliability separately among poor (score 0 – 4), intermediate (score 5 – 7), and good (score 8 – 9) preparations. ▶ **Fig. 3** demonstrates, arithmetically, that the lowest reliability was among intermediate preparation ratings, although with wide, overlapping 95 %CIs.



▶ **Fig. 2** Proportion of Boston Bowel Preparation Scale score according to the three raters.



▶ **Fig. 3** Inter-rater reliability categorized by preparation quality.

Finally, inter-rater reliability for preparations adequate and inadequate to detect 5 mm lesions are presented in ▶ **Appendix 3**.

Discussion

All three commonly used contemporary scales studied [9 – 15] have been partially validated, though head-to-head comparisons are lacking, which justifies this new assessment [16].

Scales evaluating amount of fluid present, or need for washing or suctioning (e. g. Ottawa Bowel Preparation Scale) were not included, as they are used more for research purposes [9, 16].

The BBPS showed excellent intra-rater reliability, though inter-rater reliability varied widely from fair to excellent with an overall fair-to-good inter-rater reliability. No statistically significant association was found between scores and adenoma detection. However, higher scores (mean 6.1 – 7.1) were significantly associated with the ability to detect lesions ≥ 5 mm compared with preparations judged inadequate to do so (mean 4.5 – 5.1). The BBPS was also considered the easiest scale to use.

Reliability of the CBPS demonstrated excellent intra-rater reliability and good-to-excellent inter-rater reliability. Overall inter-rater reliability was excellent. Mean CBPS scores were higher among patients in whom adenomas were detected, though these results lacked statistical significance. Higher scores were also significantly associated with ability to detect lesions ≥ 5 mm (mean 28.5 – 30.8 vs. 20.9 – 23.2). The application of this scale, however, was considered to be more complex than that of the BBPS.

The HCS did not perform as well as its counterparts in reliability testing. Though intra-rater reliability was substantial, inter-rater reliability was only fair to moderate with overall inter-rater reliability being poor. Of note, the categorical nature of the HCS may have negatively affected inter- and intra-rater reliability when compared with scales such as the BBPS and CBPS, which are graded continuously. Indeed, though the difference between scores for a given preparation may be small when graded along a continuous scale, these scores may fall into different categories when a cutoff is chosen, dividing ratings into mutually exclusive categories such as adequate and inadequate. For this reason, variance of scoring may be narrower when using a continuous scale compared with a binomial scale. No significant association was found between adequate cleansing and adenoma detection. Preparations judged adequate to detect lesions ≥ 5 mm were associated with a higher proportion of adequate HCS scores (90.3 % – 98.2 % adequate) than those judged inadequate to do so (42.0 % – 72.0 % adequate). The HCS is more complex and therefore difficult to apply in clinical practice, although it may be appropriate for research purposes.

Based on these data, BBPS and CBPS are the most reliable and most clinically relevant to routine practice, with the BBPS being considered the easiest to use. Further validation for clinical end points, including detection of lesions ≥ 5 mm, is required. This is in keeping with the results of a recent systematic review of validated scales for colon cleansing [17].

Our exploratory analysis suggests that some raters have a tendency to rate bowel preparations more strictly than others, thereby affecting inter-rater reliability. This highlights the need for calibration. In this study, to avoid favoring one scale above the rest, the BBPS training video was not used. However, prior to the study, raters were provided with clear instructions on the use of each scale as well as how to interpret the video recordings, and a calibration image was used to illustrate the presence of a 5 mm lesion in the colon. Furthermore, the lack of formal calibration in this study may better reflect reality, thus increasing generalizability of the results, as we suspect not all endoscopists in routine practice will take the time to complete available calibration exercises. The impact of calibration on inter-rater reliability requires further study.

In addition, for the BBPS, reliability appeared possibly more heterogeneous in exploratory analyses among preparations with intermediate scores (5 – 7), with significant important clinical implications in determining intervals to repeat colonoscopy. The addition of a secondary criterion (all segments need a score of at least 2) improves inter-rater reliability but may be more difficult to capture in personal or programmatic auditing practices.

The strengths of our study include the randomized order in which each of the three scales was completed by each of the raters, face-to-face comparison of the reliability of the three scales, and assessment of both intra-rater and inter-rater reliability using standardized colonoscopy videos.

This study is also unique in that raters were blinded to clinical outcomes. This is in contrast to previous studies in which endoscopists provided a rating of the bowel preparation while performing the procedure and commenting on end points, such as endoscopic findings and planned follow-up; our study was therefore perhaps less prone to the potential bias this may introduce.

Unfortunately, we were limited by local resources in the number of raters, which restricted the power of the intra-rater analysis, although this number is in keeping with the standard upheld by previously published validation studies. An effort was made to use raters of different levels of expertise, as well

as videos with varying levels of cleanliness including a large proportion of preparations of intermediate cleanliness. Though the junior trainees underwent formal training in the use of the bowel preparation scales and had extensive exposure to endoscopic procedures, it is possible that assessment by trainees may overestimate reliability, as these raters may have a tendency to adhere more rigorously to bowel preparation scale definitions than senior endoscopists who have developed certain patterns over years of experience. However, an ongoing study comparing junior and senior endoscopists suggests that assessment of adequacy of bowel preparation to detect 5 mm lesions may be made by physicians of different levels of experience (Barkun A. N., personal communication). Furthermore, the inclusion of trainees in this study may favor generalizability, as bowel preparation scales may be used by endoscopists of varying levels of experience.

The lack of statistically significant associations in our validity testing may have been due in part to low adenoma detection rates. Indeed, 72.3 % of these patients' colonoscopies were performed for cancer screening or surveillance, and many of these patients had already undergone previous colonoscopies. This heterogeneity among the study population likely translated into variable inherent risks of having an adenoma at the time of the study. Furthermore, other patient characteristics, such as smoking, male sex, and age, may have also have hindered statistical significance, as these factors have been correlated with both incidence of adenomas and suboptimal bowel preparation [16, 24].

The US Multi-Society Task Force recommendation that a bowel preparation be considered adequate if it allows for the detection of lesions greater than 5 mm has yet to be fully incorporated into any of the existing bowel preparation scales.

Conclusion

In summary, intra-rater reliability was greater than inter-rater reliability. These were lowest for the HCS, though this may be due in part to the categorical nature of this scale. All scales discriminated significantly with regard to the ability to detect lesions ≥ 5 mm, though no statistically significant association was made with adenoma detection. Overall, a review of the different test performance characteristics suggests that the BBPS and CBPS are the most discriminant and most clinically relevant for routine practice, with the BBPS being considered the easiest to use. Whereas adequacy thresholds have been suggested in the literature for the BBPS, there are few data available for the CBPS and further studies are needed to determine how to best operationalize this scale. Therefore, further validation is needed to identify the optimal scale with an emphasis on determining an adequate threshold to predict clinical outcomes and to guide clinicians in determining the appropriate interval to the next colonoscopy.

Competing interests

A. Barkun is a consultant for Pendopharm Inc., Boston Scientific Inc., Olympus Canada Inc., and Cook Inc. He has also received "at arms-length" grant funding from Boston Scientific Inc., Cook Inc. and Pendopharm. V. Heron, R. Parmar, C. Ménard, and M. Martel have no relevant conflicts of interest.

References

[1] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. CA Cancer J Clin 2015; 65: 5 – 29

[2] Centers for Disease Control and Prevention. Vital signs: colorectal cancer screening, incidence, and mortality – United States, 2002–2010. MMWR Morb Mortal Wkly Rep 2011; 60: 884 – 889

[3] Edwards BK, Ward E, Kohler BA et al. Annual report to the nation on the status of cancer, 1975–2006, featuring colorectal cancer trends and impact of interventions (risk factors, screening, and treatment) to reduce future rates. Cancer 2010; 116: 544 – 573

[4] Yang DX, Gross CP, Soulos PR et al. Estimating the magnitude of colorectal cancers prevented during the era of screening: 1976 to 2009. Cancer 2014; 120: 2893 – 2901

[5] Harewood GC, Sharma VK, de Garmo P. Impact of colonoscopy preparation quality on detection of suspected colonic neoplasia. Gastrointest Endosc 2003; 58: 76 – 79

[6] Froehlich F, Wietlisbach V, Gonvers JJ et al. Impact of colonic cleansing on quality and diagnostic yield of colonoscopy: the European Panel of Appropriateness of Gastrointestinal Endoscopy European multicenter study. Gastrointest Endosc 2005; 61: 378 – 384

[7] Johnson DA, Barkun AN, Cohen LB et al. Optimizing adequacy of bowel cleansing for colonoscopy: recommendations from the US Multi-Society Task Force on colorectal cancer. Gastroenterology 2014; 147: 903 – 924

[8] Aronchick C, Lipshultz W, Wright S. Validation of an instrument to assess colon cleansing. Am J Gastroenterol 1999; 94: 2667

[9] Rostom A, Jolicoeur E. Validation of a new scale for the assessment of bowel preparation quality. Gastrointest Endosc 2004; 59: 482 – 486

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

[11] 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

[12] Gao Y, Lin JS, Zhang HD et al. Pilot validation of the Boston Bowel Preparation Scale in China. Dig Endosc 2013; 25: 167 – 173

[13] Schindler AE, Chan WW, Obstein KL. Reliability of the Boston Bowel Preparation Scale in the endoscopy nurse population. Gastrointest Endosc 2012; 75: AB298

[14] Halphen M, Heresbach D, Gruss HJ et al. Validation of the Harefield Cleansing Scale: a tool for the evaluation of bowel cleansing quality in both research and clinical practice. Gastrointest Endosc 2013; 78: 121 – 131

[15] Gerard DP, Foster DB, Raiser MW et al. Validation of a new bowel preparation scale for measuring colon cleansing for colonoscopy: the Chicago Bowel Preparation Scale. Clin Transl Gastroenterol 2013; 4: e43

[16] Parmar R, Martel M, Rostom A et al. Validated scales for colon cleansing: a systematic review. Am J Gastroenterol 2016; 111: 197 – 204

[17] Clark BT, Rustagi T, Laine L. What level of bowel prep quality requires early repeat colonoscopy: systematic review and meta-analysis of

the impact of preparation quality on adenoma detection rate. Am J Gastroenterol 2014; 109: 1714 – 1723

[18] Leddin D, Enns R, Hilsden R et al. Colorectal cancer surveillance after index colonoscopy: guidance from the Canadian Association of Gastroenterology. Can J Gastroenterol 2013; 27: 224 – 228

[19] Calderwood AH, Schroy PC 3rd, Lieberman DA et al. Boston Bowel Preparation Scale scores provide a standardized definition of adequate for describing bowel cleanliness. Gastrointest Endosc 2014; 80: 269 – 276

[20] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159 – 174

[21] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979; 86: 420 – 428

[22] Fleiss J. Design and analysis of clinical experiments. New York: Wiley Classic Library1999: 1 – 32

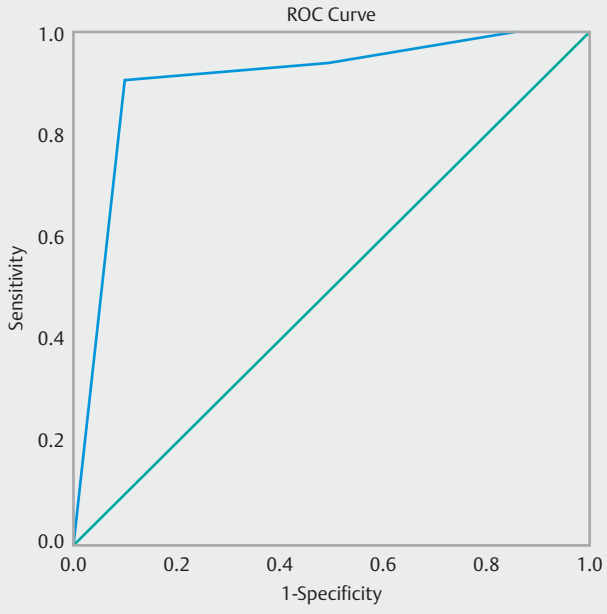
[23] Giraudeau B, Mary JY. Planning a reproducibility study: how many subjects and how many replicates per subject for an expected width of the 95 per cent confidence interval of the intraclass correlation coefficient. Stat Med 2001; 20: 3205 – 3214

[24] Wong MC, Ching JY, Chan VC et al. Determinants of bowel preparation quality and its association with adenoma detection: a prospective colonoscopy study. Medicine 2016; 95: e2251

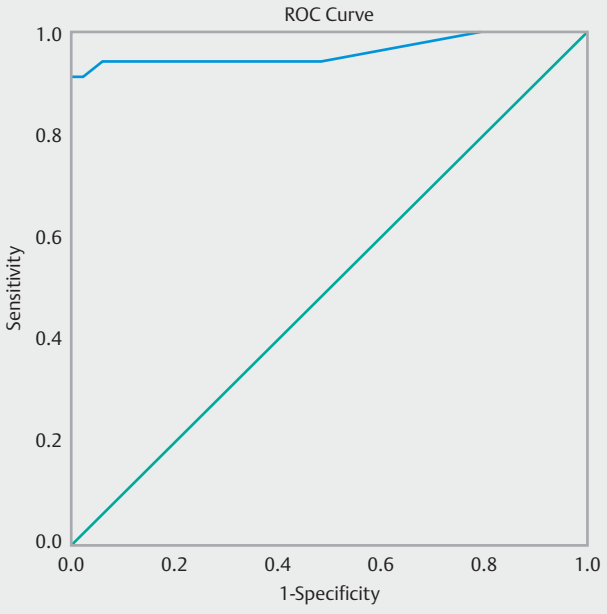
▶ **Appendix1** Scores associated with withdrawal times and with recommended interval to next colonoscopy.

	Scores associated with withdrawal time			Scores associated with recommended interval for the next colonoscopy		
	≤7 minutes	>7 minutes	*P* value	≤1 year (n=2)	>1 year (n=57)	*P* values
BBPS[1], mean ± SD						
▪ Rater 1	5.1 ± 1.6	5.2 ± 1.5	0.77	3.0 ± 2.8	5.3 ± 1.3	0.46
▪ Rater 2	6.5 ± 1.5	6.5 ± 1.5	0.91	5.0 ± 2.8	6.7 ± 1.2	0.06
▪ Rater 3	5.1 ±1.4	5.1 ± 1.1	0.81	4.0 ± 1.4	5.1 ± 1.2	0.19
CBPS%, mean ± SD						
▪ Rater 1	23.5 ± 6.1	23.9 ± 6.0	0.73	15.0 ± 14.1	24.6 ± 4.6	0.01
▪ Rater 2	28.2 ± 5.2	28.6 ± 5.7	0.75	23.0 ± 11.3	29.4 ± 3.9	0.04
▪ Rater3	24.4 ± 6.9	25.0 ± 4.9	0.81	17.5 ± 10.6	24.9 ± 5.6	0.08
HCS[2], % (95 %CI)						
▪ Rater 1	86.1 (74.2 to 98.0)	68.1 (54.3 to 81.9)	0.06	50.0 (−58.5 to 68.5)	76.5 (67.1 to 86.0)	0.34
▪ Rater 2	94.3 (86.2 to 102.4)	87.0 (76.8 to 97.1)	0.46	100.0	94.6 (88.4 to 100.7)	0.99
▪ Rater 3	72.2 (56.9 to 57.6)	55.3 (40.6 to 70.1)	0.11	50.0 (−58.5 to 68.5)	68.4 (56.0 to 80.9)	0.54

BBPS, Boston Bowel Preparation Scale; CI, confidence interval; CBPS, Chicago Bowel Preparation Scale; HCS, Harefield Cleansing Scale.
[1] Total scores.
[2] Scores A or B.



a Diagonal segments are produced by ties.



b Diagonal segments are produced by ties.

▶ **Appendix 2** Receiver operating characteristic curve for ability to detect 5 mm lesions based on bowel cleanliness. **a** Boston Bowel Preparation Scale (AUC: 0.91 [0.83 to 0.98]) Threshold: Score 6; sensitivity 90.9 % (75.7 % to 98.1 %); specificity 90 % (78.2 %; 96.7 %) **b** Chicago Bowel Preparation Scale (AUC 0.96 [0.91 to 1.00]) Threshold: Score 26; sensitivity 93.9 % (79.8 % to 99.3 %); specificity 88.0 % (76.7 % to 95.5 %).

▶ **Appendix 3** Overall inter-rater reliability among preparations adequate to detect 5 mm lesions vs. inadequate to do so.

	Adequate to detect 5 mm lesions	Inadequate to detect 5 mm lesions
BBPS[1], ICC, % (95 %CI)	0.63 (0.28 to 0.81)	0.64 (0.22 to 0.82)
CBPS[1], ICC, % (95 %CI)	0.43 (0.50 to 0.69)	0.70 (0.38 to 0.85)
HCS[2], κ (95 %CI)	0.16 (– 0.05 to 0.35)	0.31 (0.15 to 0.47)

ICC, interclass correlation coefficient; CI, confidence interval; BBPS, Boston Bowel Preparation Scale; CBPS, Chicago Bowel Preparation Scale; HCS, Harefield Cleansing Scale.
[1] Total scores.
[2] Scores A or B.