A proposed staging system and stage-specific interventions for familial adenomatous polyposis

Patrick M. Lynch, MD,[1] Jeffrey S. Morris, PhD,[2] Sijin Wen, PhD,[3] Shailesh M. Advani, MD, MPH,[1] William Ross, MD,[1] George J. Chang, MD, MS,[4] Miguel Rodriguez-Bigas, MD, FACS, FASCRS,[4] Gottumukkala S. Raju, MD, FASGE,[1] Luigi Ricciardiello, MD,[5] Takeo Iwama, MD,[6] Benedito M. Rossi, MD, PhD,[7] Maria Pellise, MD, PhD,[8] Elena Stoffel, MD, MPH,[9] Paul E. Wise, MD,[10] Lucio Bertario, MD,[11] Brian Saunders, MD, PRCP,[12] Randall Burt, MD,[13] Andrea Belluzzi, MD,[14] Dennis Ahnen, MD,[15] Nagahide Matsubara, MD,[16] Steffen Bülow, MD, DMSc,[17] Niels Jespersen, MD,[17] Susan K. Clark, MD, FRCS,[18] Steven H. Erdman, MD,[19] Arnold J. Markowitz, MD,[20] Inge Bernstein, MD, PhD, MHM,[21] Niels De Haas, MD,[21] Sapna Syngal, MD, MPH,[22] Gabriela Moeslein, MD[23]

Houston, Texas; Morgantown, West Virginia; Ann Arbor, Michigan; St. Louis, Missouri; Salt Lake City, Utah; Denver, Colorado; Columbus, Ohio; New York, New York; Boston, Massachusetts, USA; Milan, Bologna, Italy; Middlesex, United Kingdom; Saitama, Hyogo, Japan; Sao Paulo, Brazil; Barcelona, Spain; Copenhagen, Aalborg, Denmark; Bochum, Germany

Background and Aims: It is not possible to accurately count adenomas in many patients with familial adenomatous polyposis (FAP). Nevertheless, polyp counts are critical in evaluating each patient's response to interventions. However, the U.S. Food and Drug Administration no longer recognizes the decrease in polyp burden as a sufficient chemoprevention trial treatment endpoint requiring a measure of "clinical benefit." To develop endpoints for future industry-sponsored chemopreventive trials, the International Society for Gastrointestinal Hereditary Tumors (InSIGHT) developed an FAP staging and intervention classification scheme for lower-GI tract polyposis.

Methods: Twenty-four colonoscopy or sigmoidoscopy videos were reviewed by 26 clinicians familiar with diagnosis and treatment of FAP. The reviewers independently assigned a stage to a case by using the proposed system and chose a stage-specific intervention for each case. Our endpoint was the degree of concordance among reviewers staging and intervention assessments.

Results: The staging and intervention ratings of the 26 reviewers were highly concordant ($\rho = 0.710$; 95% credible interval, 0.651-0.759). Sixty-two percent of reviewers agreed on the FAP stage, and 90% of scores were within ±1 stage of the mode. Sixty percent of reviewers agreed on the intervention, and 86% chose an intervention within ±1 level of the mode.

Conclusions: The proposed FAP colon polyposis staging system and stage-specific intervention are based on a high degree of agreement on the part of experts in the review of individual cases of polyposis. Therefore, reliable and clinically relevant means for measuring trial outcomes can be developed. Outlier cases showing wide scatter in stage assignment call for individualized attention and may be inappropriate for enrollment in clinical trials for this reason. (Gastrointest Endosc 2016;84:115-25.)

Abbreviations: FAP, familial adenomatous polyposis; FDA, Food and Drug Administration; InSiGHT, International Society for Gastrointestinal Hereditary Tumors; ICC, intraclass correlation coefficient; IPSS, InSiGHT polyposis staging system.

DISCLOSURE: *Dr Burt is a consultant for Myriad Genetics. Each video reviewer received compensation for the time devoted to review of endoscopic videos and for completing the scoring forms. All other authors disclosed no financial relationships relevant to this publication. Support for this study was provided by SLA Pharma (UK) Ltd. The protocol number for this study at MD Anderson Cancer Center is PA11-0926.*

0016-5107/$36.00
http://dx.doi.org/10.1016/j.gie.2015.12.029

Received July 16, 2015. Accepted December 28, 2015.

Current affiliations: Department of Gastroenterology, Hepatology, and Nutrition, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA (1), Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA (2), Department of Biostatistics, West Virginia University Health Sciences Center, Morgantown, West Virginia, USA (3), Department of Surgical Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA

(footnotes continued on last page of article)

It is virtually impossible to accurately count adenomas during endoscopy in many patients with familial adenomatous polyposis (FAP). Nevertheless, polyp counts are critical in evaluating a patient's response to chemopreventive agents. However, there has been virtually no guidance for endoscopists and surgeons in determining when surgery should be performed. More pointedly, the determination of the U.S. Food and Drug Administration (FDA) that approval of new chemopreventive agents must meet a higher standard of clinical benefit has left the FAP community speculating as to what such a standard really calls for. Members of the International Society for Gastrointestinal Hereditary Tumors (InSiGHT) undertook the described study in order to develop a staging and staged intervention system that would provide an acceptable measure of clinical benefit in future industry-sponsored chemoprevention trials and other interventions in FAP.

In 1989, Spigelman et al[1] proposed a staging system for duodenal adenomas in patients with FAP. This system has enabled clinicians to monitor patients more effectively and has guided clinical interventions. Unfortunately, no corresponding staging system exists for adenomas in the colon and rectum in either the pre- or postoperative setting, perhaps because some perform colectomy or proctocolectomy soon after diagnosis of colorectal adenomas, regardless of severity. But many clinicians use the extent of "polyp burden" and clinical judgment to determine the timing of colectomy, both of which are subjective and individual based, thus indicating a need for standardization.

A diagnosis of FAP is typically established on the basis of adenomatous polyposis coli gene testing, and adenomas can be found in patients as young as age 10 or 12.[2,3] Although it is a normal practice to operate at an early point in the evolution of FAP, there has been a tendency to defer surgery in these young patients. Improvements in endoscopes and better, safer anesthesia for pediatric use have made full colonoscopy a very acceptable procedure in children. There is also value in waiting for the rectum to "declare itself" insofar as the development of adenoma burden is concerned, so that surgeons can better select the appropriate operation: colectomy or proctocolectomy.[4] Conversely, much older patients with attenuated FAP and mutY homolog (*MUTY*)-associated polyposis may initially be diagnosed with a very mild adenoma burden at age 50 or later.[5,6] An unknown but small fraction of such patients can be managed conservatively, with periodic multiple polypectomies without surgery.

This emerging diversity in FAP presentation, diagnosis, and treatment has not, of itself, been enough to stimulate the development of a colorectal polyposis staging system. However, in 2011, in a letter, the FDA stated that it would no longer approve, much less accelerate approval of, chemopreventive agents for the treatment of premalignant conditions such as FAP on the basis of a reduction in polyp number and size alone; a clearer demonstration of clinical benefit would be required (E L. Memorandum of meeting minutes pre-IND/pre-NDA for eicosapentaenoic acid [free fatty acid] [EPA-FFA]. In: Services DoHH, editor, Q8 2011:1-20).[7] In 2011, Meyskens and colleagues highlighted the need to develop effective biomarkers and true clinical endpoints for cancer chemoprevention trials.[8] At the 2011 meeting of InSiGHT, a group of FAP experts met with pharmaceutical leaders interested in responding to the FDA's clinical benefit challenge. The experts agreed that demonstrating clinical benefit would require the development of clinically relevant signposts of FAP progression that would also serve as primary endpoints for clinical trials of chemopreventive therapies. Also, treatment response or progression would have to be couched in oncological meaningful terms, despite the fact that FAP-related mortality is uncommon in patients with FAP because of current intensive endoscopic surveillance and surgical prophylaxis. To be clinically meaningful, the progressive disease stage would need to be linked to progressively more aggressive interventions. A staging system for colorectal polyposis akin to the Spigelman et al[1] staging system for duodenal polyposis might thus provide objective and clinically relevant measures of time to disease progression as well as disease regression. As a subgroup of the FAP experts who met in 2011, we undertook the development and testing of such a staging system.

As detailed in the following, we created a scale that divides colorectal polyposis into 5 progressive stages based on adenoma number and size. The degree of dysplasia, age, and desmoid disease were not considered in developing the InSiGHT polyposis staging system (IPSS). We then created a corresponding scale specifying the endoscopic, surgical, and/or chemopreventive interventions considered appropriate to the adenoma burden. Recognizing that clinical staging and interventions are based on expert opinion, we convened a panel of experts—endoscopists and surgeons—to review videos of edited colonoscopies or sigmoidoscopies (in cases of prior colectomy or proctocolectomy). Our endpoint was to discern the degree of agreement among the experts in assigning a given video to one of the 5 predefined InSiGHT polyposis staging system (IPSS) stages and, further, in proposing appropriate interventions for the stages they assigned.

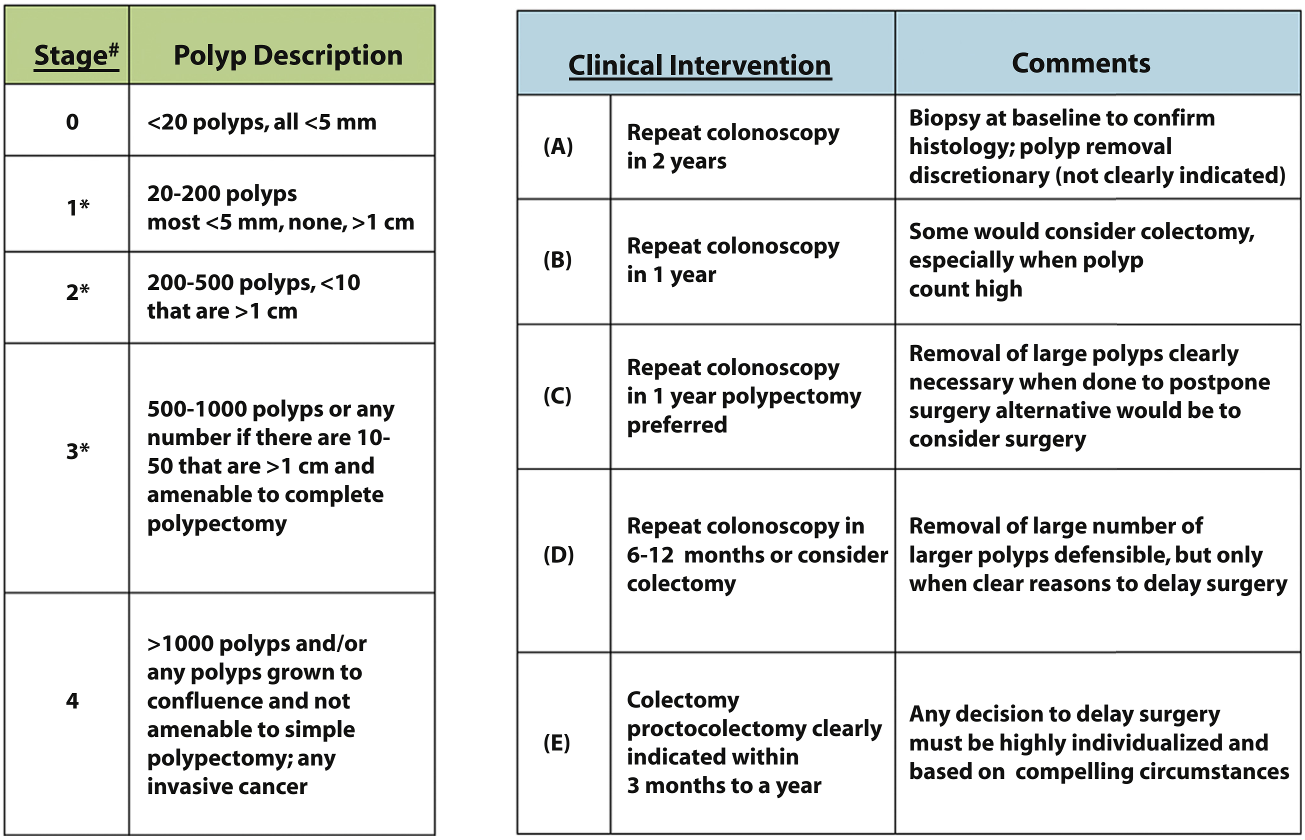
METHODS

Development of the IPSS. At the 2011 annual InSiGHT meeting, the need for a staging system for colorectal polyposis was recognized in response to the FDA position requiring a measure of clinical benefit for new drug approval. Therefore, we developed an arbitrary classification system for progressive categories of colorectal polyposis severity and a means for validating that classification. The categories were developed by the primary author (P.M.L.) with the expectation that a given range of severity should lend itself to interventions appropriate to that degree of severity. Delay in progression from 1 stage to a higher stage or regression to a lower stage should translate into a change in necessary intervention and thus constitute a worthwhile measure of clinical benefit. The proposed classification is seen in Figures 1 and 2. The initial test of suitability of this staging system was to be based on a review of a large number of videos of FAP cases by a large panel of clinicians, most of whom are recognized experts in FAP management. The system was developed to represent the broad grouping of polyp burden (both in number and size) in such a way that one could assign a given case to a broad category with reasonable confidence but without the need to undertake an attempt to accurately count the polyps. A 5-point numbered scale (0-4) was used for the staging system, and a 5-point letter scale (A-E) was used for classifying stage-specific interventions. The system was developed so that the intervention system (A-E) would correspond to the stage identified. However, the reviewers were not notified about this classification to prevent any biases or direct them to a specific intervention.

Data collection

We contacted IPSS members who are experts in the field of FAP. These members were e-mailed a detailed description of the study and were requested to respond via e-mail regarding their interest in participation in the study. Participants who agreed to participate formed our list of reviewers. Participation in the study implied informed consent. Because this study did not pose any harm and/or risk to the reviewers, no signed informed consent was needed. A total of 29 experts agreed to participate in the review of 24 videos; 26 (90%) completed the study. The study was approved by the Institutional Review Board at The University of Texas MD Anderson Cancer Center with a waiver of consent for use of the previously obtained and deidentified videos that comprised the study material.

We collected archived and deidentified videos of colonoscopies and sigmoidoscopies performed during earlier multicenter chemoprevention trials conducted at the MD Anderson Cancer Center, the Cleveland Clinic, and St. Mark's Hospital.[9] One of the authors (P.M.L.) selected 24 videos from the archive to represent a range of FAP severity. These videos represent the typical distribution of FAP cases that we experience in clinical settings. They were taken

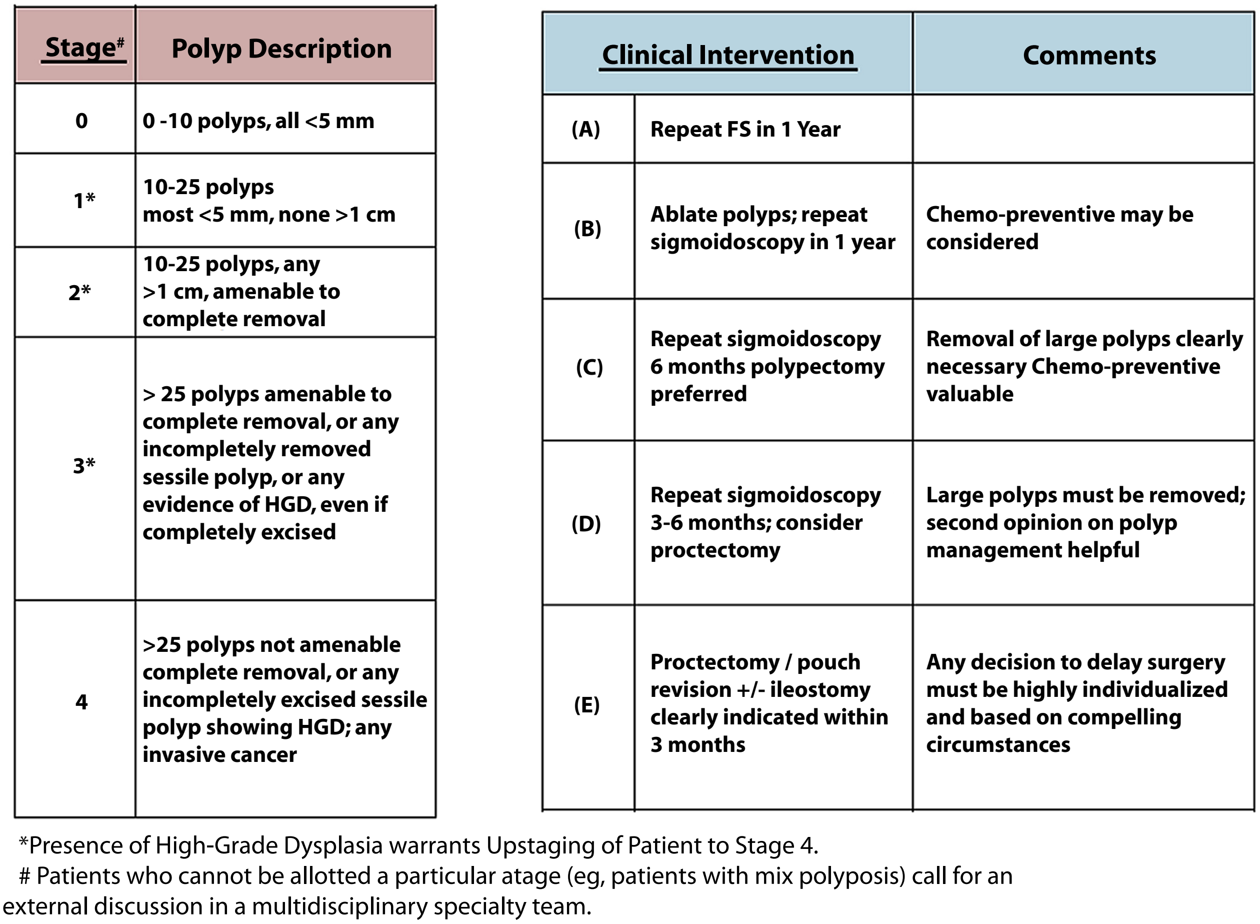


Stage#	Polyp Description
0	<20 polyps, all <5 mm
1*	20-200 polyps most <5 mm, none, >1 cm
2*	200-500 polyps, <10 that are >1 cm
3*	500-1000 polyps or any number if there are 10-50 that are >1 cm and amenable to complete polypectomy
4	>1000 polyps and/or any polyps grown to confluence and not amenable to simple polypectomy; any invasive cancer

Clinical Intervention		Comments
(A)	Repeat colonoscopy in 2 years	Biopsy at baseline to confirm histology; polyp removal discretionary (not clearly indicated)
(B)	Repeat colonoscopy in 1 year	Some would consider colectomy, especially when polyp count high
(C)	Repeat colonoscopy in 1 year polypectomy preferred	Removal of large polyps clearly necessary when done to postpone surgery alternative would be to consider surgery
(D)	Repeat colonoscopy in 6-12 months or consider colectomy	Removal of large number of larger polyps defensible, but only when clear reasons to delay surgery
(E)	Colectomy proctocolectomy clearly indicated within 3 months to a year	Any decision to delay surgery must be highly individualized and based on compelling circumstances

*Presence of High-Grade Dysplasia Warrants Upstaging of Patient to Stage 4.
# Patients who cannot be allotted a particular stage (eg, patients with mix polyposis) call for an external discussion is a multidisciplinary specialty team.

Figure 1. Proposed InSiGHT staging system classification and clinical interventions for colonic polyposis.



Stage#	Polyp Description
0	0 -10 polyps, all <5 mm
1*	10-25 polyps most <5 mm, none >1 cm
2*	10-25 polyps, any >1 cm, amenable to complete removal
3*	> 25 polyps amenable to complete removal, or any incompletely removed sessile polyp, or any evidence of HGD, even if completely excised
4	>25 polyps not amenable complete removal, or any incompletely excised sessile polyp showing HGD; any invasive cancer

Clinical Intervention		Comments
(A)	Repeat FS in 1 Year	
(B)	Ablate polyps; repeat sigmoidoscopy in 1 year	Chemo-preventive may be considered
(C)	Repeat sigmoidoscopy 6 months polypectomy preferred	Removal of large polyps clearly necessary Chemo-preventive valuable
(D)	Repeat sigmoidoscopy 3-6 months; consider proctectomy	Large polyps must be removed; second opinion on polyp management helpful
(E)	Proctectomy / pouch revision +/- ileostomy clearly indicated within 3 months	Any decision to delay surgery must be highly individualized and based on compelling circumstances

*Presence of High-Grade Dysplasia warrants Upstaging of Patient to Stage 4.

# Patients who cannot be allotted a particular atage (eg, patients with mix polyposis) call for an external discussion in a multidisciplinary specialty team.

Figure 2. Proposed InSiGHT staging system classification and clinical interventions for postcolectomy cases with ileorectal anastomosis. *FS*, flexible sigmoidoscopy; *HGD*, high-grade dysplasia.

from patients who met the criteria for FAP and who had participated in previous chemoprevention trials.[10] The videos were loaded into a video editing program (Corel Video Studio ProX7, Corel Corporation, Ottawa, Canada) and edited to capture total adenoma burden while preventing reviewer fatigue by eliminating extraneous footage (all videos ran <2 minutes). Consequently, deidentified and sequentially numbered videos were transferred to USB thumb drives and mailed to reviewers. The USB drives also included an instruction page with a link to the data-recording site in Survey Monkey. Raters were provided with tables displaying the proposed IPSS guidelines for staging FAP of the colon or FAP of the rectum only (for postcolectomy patients), and the proposed stage-specific interventions, which were arbitrarily chosen for the purposes of this study (Figs. 1 and 2).

The reviewers were InSiGHT members known to be experienced with FAP, and other institutional colleagues recommended by these members. Demographic characteristics of the reviewers are summarized in Table 1. Reviewers received nominal reimbursement for their participation in video review and scoring process. In addition to assigning a stage for and choosing a recommended level of intervention for each FAP case depicted in the videos, reviewers were asked to provide comments after scoring each video. Reviewers were also required to self-designate themselves as either surgeons or endoscopists and record their annual volume of patients with FAP. Having scored the videos and assigned a recommended intervention, the reviewers were then asked to rate the utility of the IPSS and of the stage-specific interventions by using a 5-point visual analog scale ranging from “strongly agree” to “strongly disagree.”

Statistical design

For each video, we provided reviewers with electronic scoring sheets consisting of 2 ordinal 5-point scales as discussed previously, and our goal was to assess multirater concordance based on these ordinal scales. Because there are not any standard measures of concordance that apply to our setting with an ordinal (5-point) scale and multiple raters, we used a Bayesian multiple-rater model to assess the concordance of the ordinal data across raters.[11,12] This method allowed us to estimate the rater variation

TABLE 1. Summary of characteristics of 26 reviewers for the IPSS staging system (N = 26)

Characteristic	No. (%)
Sex	
Male	20 (77)
Female	6 (23)
Specialty	
Colorectal surgery	13 (50)
Gastroenterology	13 (50)
Clinical category	
Endoscopist	14 (54)
Surgeon	12 (46)
Workplace setting	
Clinical	5 (19)
Academic	21 (81)
No. of FAP patients seen yearly	
0-5	3 (11)
6-10	4 (15)
11-20	6 (24)
≥21	13 (50)

FAP, Familial adenomatous polyposis; *IPSS*, InSiGHT polyposis staging system.

and overall variation and, therefore, to obtain a model-based intraclass correlation coefficient (ICC), ρ, as our measure of concordance. The details of this measure and its calculation are provided in the Appendix (available online at www.giejournal.org). Briefly, we assumed that the variability across ratings had 2 components: a rater-to-rater variability and a video-to-video variability. The measure ρ indicates the proportion of total variability attributed to the video-to-video component and is constrained to be between 0 and 1. Thus, higher ρ indicates greater concordance, with $\rho = 1$ indicating that all raters gave the same rating to all videos and $\rho = 0.5$ indicating that the variability across raters was equal in magnitude to the variability across videos. We plotted the data in heat maps, graphic representations of tables, by using colors to represent numbers. In the heat maps (Figs. 3–5), the darker the boxes, the higher the proportion of reviewers in agreement.

The multiple-rater model was fit by using a Markov chain Monte Carlo algorithm with 10,000 samples after a burn-in of 5000 used for inference. From these samples, we computed the posterior mean, standard error, and 95% credible interval for ρ and tested the null hypothesis that $\rho \leq 0.5$ by computing $P = \text{Prob}\ (\rho \leq 0.5|\text{data})$, rejecting the null hypothesis if $P < .05$.

Because there is no standard sample size software for this multiple-rater ordinal measure of concordance, we performed simulations to determine a sample size that would provide sufficient power to detect a strong concordance. We simulated 100 trials with 24 raters and 24 videos with 5 different scenarios: (1) the rater variation is the same as the video variation ($\rho = 0.5$), (2) the rater variation is 1/2 of the video variation ($\rho = 0.67$), (3) the rater variation is 3/7 of the video variation ($\rho = 0.70$), (4) the rater variation is 1/3 of the video variation ($\rho = 0.75$), and (5) the rater variation is 1/4 of the video variation ($\rho = 0.80$). We concluded that there was significant agreement between raters if the chance of ρ being <0.5 was very small (<0.05).

Our simulation showed that a sample size of 24 raters and 24 videos would have at least 83% power to show a concordance of $\rho = 0.70$ (ie, the rater variation is 3/7 of the video variation). More results from the simulation are shown in Supplemental Tables 1 and 2 (available online at www.giejournal.org). A weighted Cohen's κ was used to assess concordance of the assigned stages and interventions for each rater.[10] R Version 2.15.2 (R Foundation, Vienna, Austria) was used to conduct statistical analysis.

RESULTS

Our study's key finding was the demonstration of strong agreement in scoring across the 24 videos by the 26 reviewers. There was high concordance across raters, with a mean ρ of 0.710 (standard error, 0.027; 95% credible interval, 0.651-0.759). From this, we rejected the null hypothesis that $\rho \leq 0.5$ ($P < .0001$). A histogram of posterior distribution of ρ is given in Supplemental Fig. 1 (available online at www.giejournal.org). We observed a statistically significant degree of agreement in both the staging of polyp burden and the selected interventions for a given stage. This agreement is important because clinical decision making in the abstract is often quite different from that applied in real cases.

Heat maps of reviewer staging by video shows that those reviewers reached a high degree of agreement at the extremes of IPSS stage. Figure 3A shows, for each video, the proportion of raters who assigned each stage. The data reveal that at the highest and lowest levels (stages 4 and 0) of adenoma burden, there was near-perfect concordance between observers in assigning a given video to a stage. Not surprisingly, perhaps, in the approximate mid-range of adenoma burden, there was greater scatter, although with an overall high level of concordance. Nonetheless, we discerned wide agreement on most videos, with most scores ranging within 1 stage "worse" or "better" than the modal value. In only 3 of the 24 videos did scores vary by more than 1 stage above or below the mode. Figures 3B and 3C present the raw ratings for each video for endoscopists and surgeons, respectively. At the extremes of severity, we found greater agreement between the surgeons and the endoscopists; the level of concordance was similar between the 2 groups.

Greater scatter was seen with respect to interventions (Fig. 4). In general, however, reviewers either agreed with

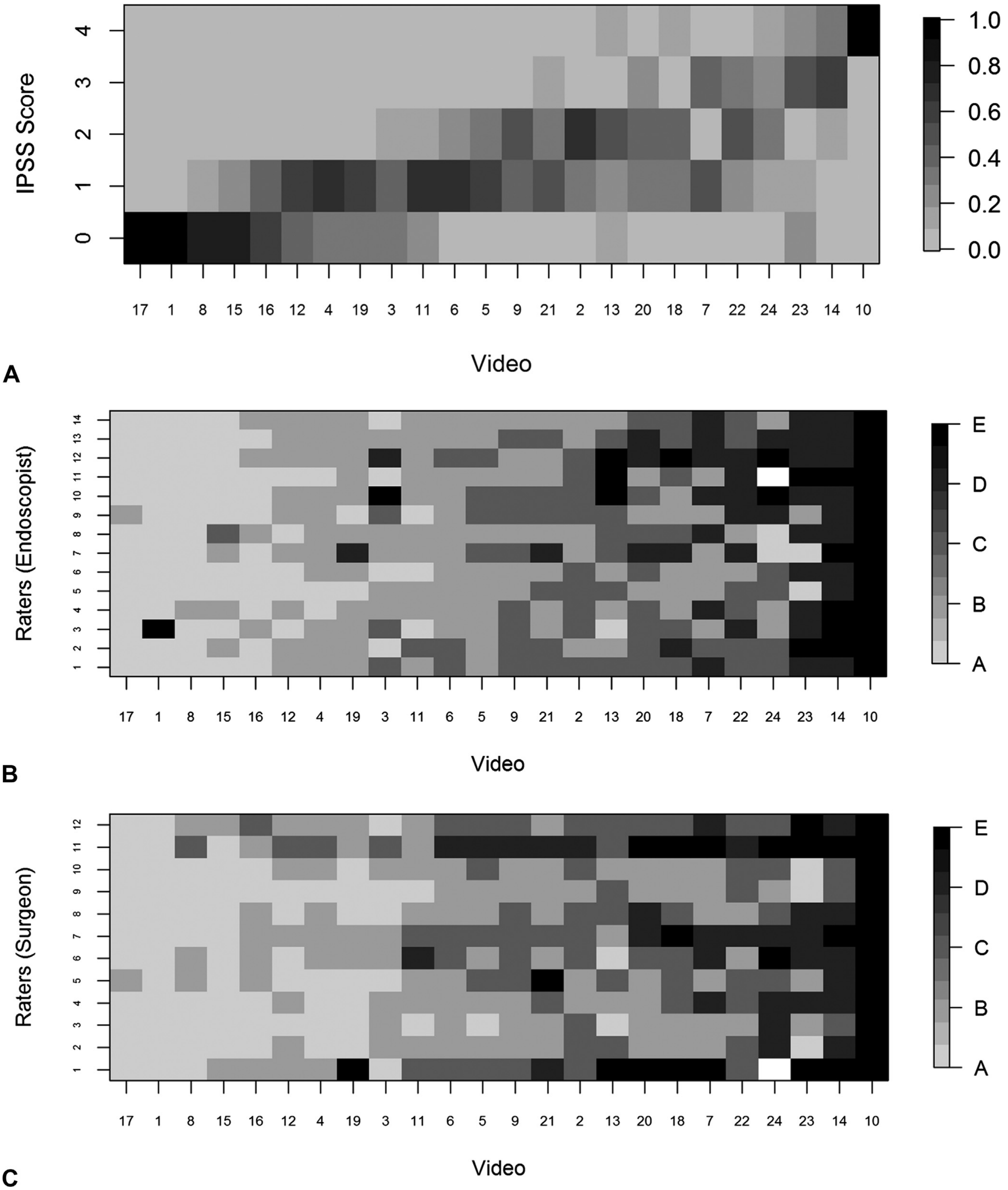


Figure 3. A, Heat map displaying the proportion of IPSS scores by video, with videos ordered from lowest average stage (video 17) to highest average stage (video 10). **B**, Heat map displaying InSiGHT polyposis staging system scores from 14 endoscopists by video. **C**, Heat map displaying InSiGHT polyposis staging system scores from 12 surgeons by video. *IPSS*, InSiGHT polyposis staging system.

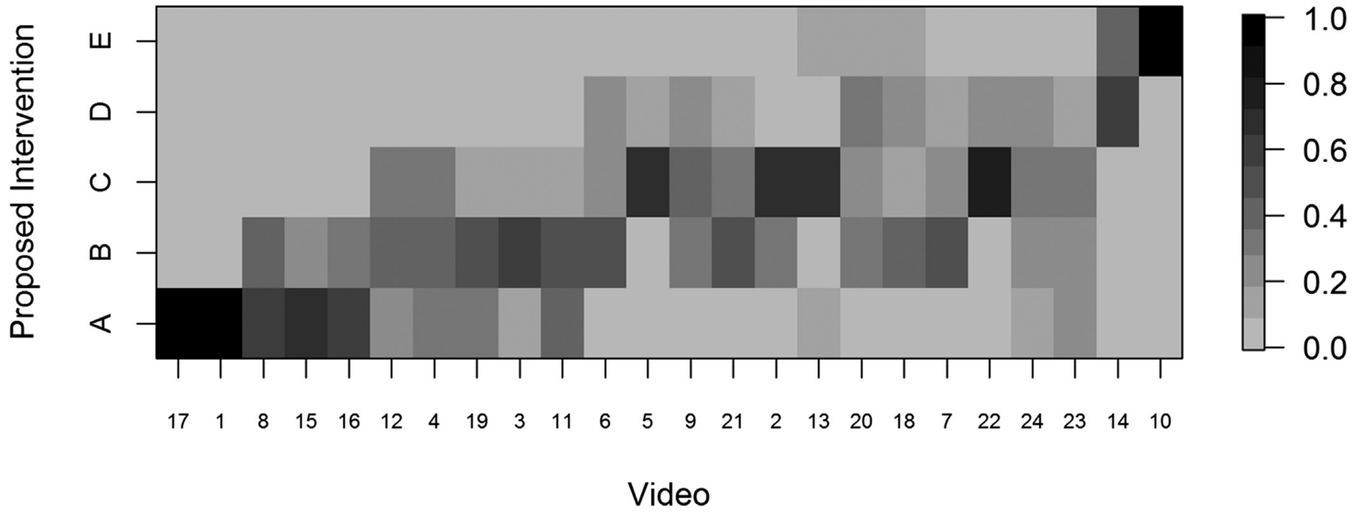


Figure 4. Heat map displaying the proportion of raters (N = 26) who assigned each intervention to each video, with rows representing reviewer-selected intervention scores ranging from A to E, and videos ordered from lowest to highest average scores.

the proposed intervention for the stage to which they had assigned a given video or recommended an intervention within 1 incremental level of the proposed one. Heat maps (Figs. 5A-D) were also constructed to demonstrate individual reviewer's tendencies to recommend a more aggressive versus less aggressive intervention relative to a particular polyp burden. We derived a score by subtracting the numeric value of each assigned stage (0-4) from the numeric value of the stage corresponding to the assigned intervention (0-4, corresponding to A-E). A difference of 0 indicated agreement with the stage-specific intervention proposed in Figure 1 or 2. A positive score indicated that the reviewer preferred a more aggressive intervention than that proposed by the researchers. Conversely, a negative score indicated that the reviewer preferred a less aggressive intervention for the polyposis burden shown. In most cases (20 of 26), the reviewers agreed with the stage-specific interventions provided by the researchers, and in 92% of cases, reviewers chose an intervention within 1 level more or less aggressive than that provided in our proposed system. We also observed a modest difference by specialty with respect to the aggressiveness of the assigned intervention. Endoscopists were slightly more likely to have a positive treatment aggressiveness score (22.1% of cases) than were surgeons (15.7% of cases), indicating endoscopists were more likely to recommend more aggressive treatment. To assess concordance of the IPSS scores and intervention, Cohen κ coefficients were computed. The κ with square weights was calculated between interventions and IPSS scores for each rater. The mean Cohen κ from 26 raters was 0.793, with a standard deviation of 0.188, demonstrating that raters predominantly tended to propose interventions coinciding with their IPSS staging for that patient. In addition, we ran a similar analysis to compare scoring based on reviewer sex and number of FAP patients seen every year (≥11 FAP patients per year vs ≤10 patients per year). There were no significant differences seen between these groups (Supplemental Table 3, available online at www.giejournal.org).

Of the 26 reviewers, 25 strongly agreed (17) and agreed (8) that "the development of a staging system for colorectal polyposis will be helpful in communicating with colleagues regarding patient status," and 1 reviewer responded "neutral" to this question. When we use the same scale, 22 of 26 reviewers, 18 strongly agreed and 4 agreed that "the development of a staging system for colorectal polyposis will be helpful in evaluating endpoints in clinical chemoprevention trials." There was also considerable support for the proposed staging system. Of 25 responses, 23 indicated that the reviewer agreed (21) or strongly agreed (2) with the proposition "[s]ubject to my specific comments in the scoring sheet above, I am in general agreement with the present proposed IPSS." When asked to rate their general agreement with the proposed stage-specific intervention scale (subject, as above, to comments offered on the scoring sheet), reviewers expressed generally supportive, although more qualified, responses. Sixteen of the 26 reviewers agreed with the proposed interventions, whereas 8 were neutral and 2 disagreed with them (see Supplemental Table 4, available online at www.giejournal.org, for these ratings).

DISCUSSION

Our study's key finding was the demonstration of strong agreement in scoring across the 24 videos by the 26 reviewers.

We developed a staging system for severity of colorectal polyposis FAP for future industry-sponsored clinical trials. We sought to determine whether experts could reach consensus as to the appropriate stage assignment and

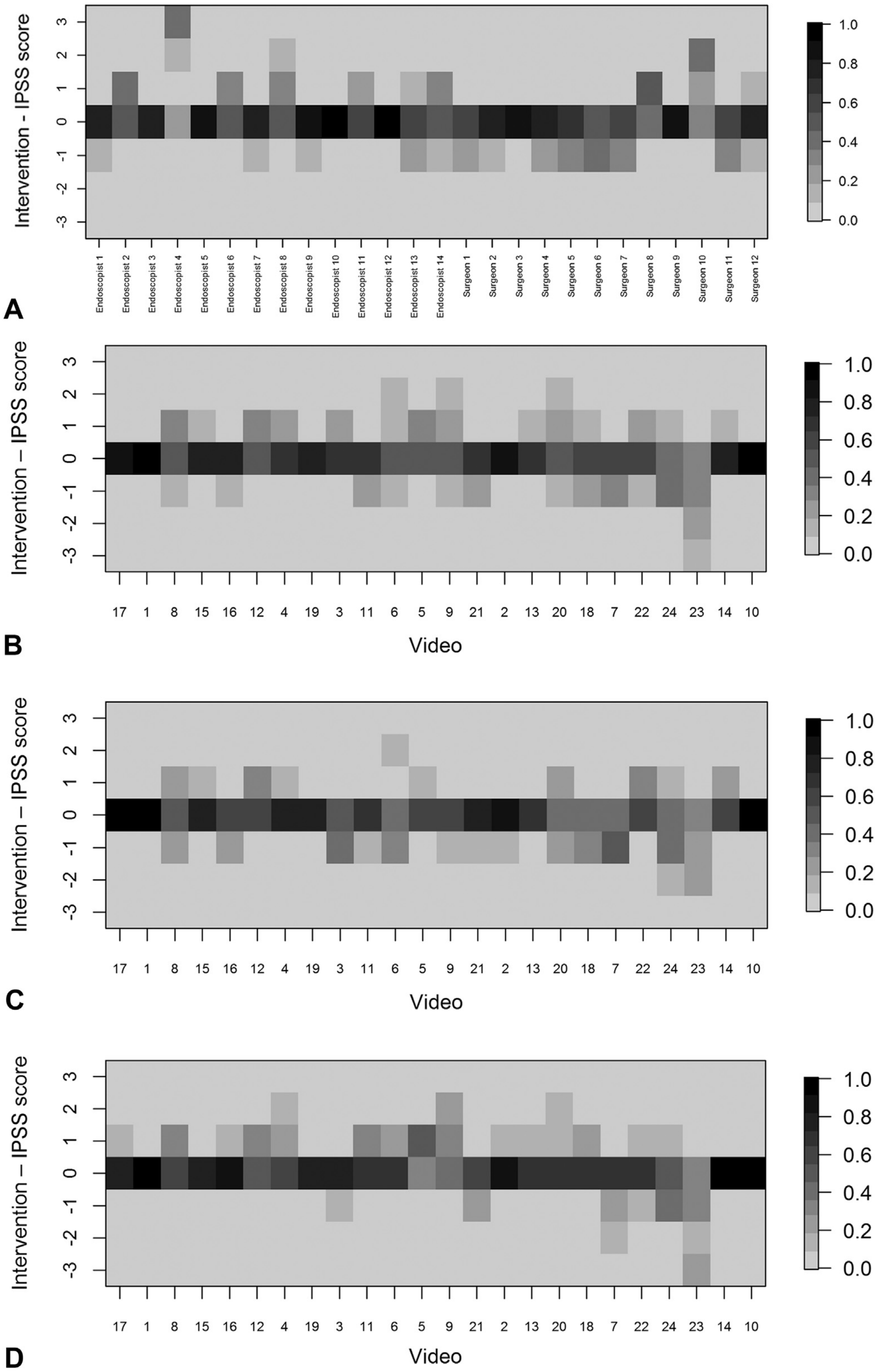


Figure 5. Heat maps displaying differences between recommended interventions and IPSS scores (intervention minus IPSS score) for each reviewer. Positive values indicate that the reviewer recommended a higher intervention level than that corresponding to the assigned stage; negative values indicate that the reviewer recommended a lower intervention level. **A**, Heat map displaying the proportion of videos with each difference value by rater. **B**, Heat map displaying proportion of raters with each difference value by video. **C**, Heat map displaying proportion of endoscopists with each difference value by video. **D**, Heat map displaying proportion of surgeons with each difference value by video. *IPSS*, InSiGHT polyposis staging.

intervention for a given case. However, surgeons and endoscopists who deal with FAP may assign the same video images to different stages. We had reason to anticipate some variability, based on our quantitative video-based study of adenoma burden, polyp number, and diameter.[13] These experts scored more than 20 colonoscopy videos by using an "electronic abacus," placing polyps into "bins" corresponding to 3 diametric ranges. The measure of scatter led us to conclude that colorectal FAP could and should be classified into broad categories if a high degree of concordance needs to be achieved. We wanted to include enough reviewers to provide more statistical clarity than was possible in our earlier study. Polyposis staging was considered the straightforward part of this exercise. We expected that observers reviewing the same endoscopic video clip would agree within the broad ranges we provided. More challenging was the proposition that an arbitrary set of interventions would be agreed on as well. Arguably, a reviewer could modify staging for a video to bring it into line with a desired intervention. In other words, reviewers might consciously or unconsciously "up-stage" or "down-stage" to arrive at a preferred intervention. Conversely, if all of a reviewer's recommended interventions were the same (eg, "needs surgery") regardless of stage, then our attention to staging would prove irrelevant. Reviewers' comments did not show a substantial amount of such cognitive dissonance or compensatory reasoning.

Our study has limitations. For one, the proposed classifications by polyp count and diameter were based on expert opinion without previous validation. True validation awaits longer term outcome measures, including the need for surgical intervention or the development of cancer. Staging categories also did not account for key factors that could affect interventions and their timing; for example, neither age nor risk of desmoid disease was factored into the IPSS. Young patients with modest polyp burdens might be managed expectantly to allow for completion of their pubescent growth spurts before colectomy, whereas surgery might be recommended straightaway for older patients with identical polyp burdens. A patient with known desmoid disease might, appropriately, choose to wait as long as possible for surgery. No feature of a colonoscopy video can properly inform clinicians on issues such as age and desmoid status, nor does the IPSS account for the degree of dysplasia in adenomas, although any adenoma with high-grade dysplasia should probably be placed in stage IV. Attempting to stratify according to such criteria would have, we think, needlessly complicated the staging system.

Because FAP is rare, its care is commonly given over to experts who have extensive experience with hereditary colorectal cancer syndromes. Determining whether a broad consensus could be reached was viewed as the first step toward establishing a staging system for colorectal polyposis. Hence, for our study, we recruited FAP reviewers who are experts in the field. Although they may share certain biases, at least such biases reflect expertise in assessing the severity of actual FAP cases. More important than the use of experts as such was the general agreement in stage and intervention assignment across the panel of reviewers. An appropriate next undertaking, but one beyond the immediate scope of this study, would be to engage trainees and other nonexperts to determine whether they could intuitively or with minimal training come to use the IPSS in a fashion similar to our experts.

We further examined results for the videos showing the greatest scatter in stage assignment. In video 13, for instance, there was 1 very large confluent adenoma in the right side of the colon, but very few other adenomas, all limited to the right side of the colon. Some scorers dismissed the confluent adenoma and assigned a low stage due to the low overall polyp count, whereas others were very influenced by the 1 large confluent adenoma and assigned a high stage. Such discordant findings were not really anticipated in the IPSS as it was initially conceived. These kinds of difficult outlying cases defy ready classification and require individual attention. It can be argued, at least, that these cases will be appropriately "flagged" by a demonstration of discordant staging on review by multiple experts. Such cases might well be excluded from clinical chemoprevention trials for this reason. (See Supplemental Table 5, available online at www.giejournal.org, for some of the reviewers' comments on 3 videos with outlier cases.) After reviewing our results, we contacted our reviewers to assess whether high-grade dysplasia would automatically classify a patient as stage IV. Fifteen of 26 respondents replied and agreed to the statement and also stated that these atypical cases require discussion and consensus in multispecialty conference settings.

Technical considerations in endoscopy can affect the performance of the IPSS or any other staging system, including colon preparation and withdrawal. There are additional measures that could reduce disagreement in staging. Cases of greatest scatter could be reviewed jointly to determine whether outliers are due to errors in the identification and classification of polyps (lymphoid polyps can be a problem), polyp counts, or polyp size. Joint review of difficult cases might also lead to modification of the staging designation when disagreement exists as to the "bins" to which such cases are assigned. Different count or dimension thresholds might lead to different criteria for staging any given polyp burden. To address these issues, tutorials similar to those used by pathologists to standardize interpretation of pathologic specimens could be designed.

Because our reviewers were FAP experts, we did not attempt to control for the amount of time spent by each reviewer to score a particular video. One might argue that less-experienced participants would tend to spend more time on difficult cases than those with more experience and perhaps reach a better score/rating than if less time were spent reviewing a case. Because we did not monitor the number of times a video was reviewed or the time spent in doing so, we cannot say what the effect of this might have been on interrater concordance. This will be addressed in future studies.

In this study, we used deidentified videos that had been used in previous chemoprevention trials. Although these represent a range of FAP cases that normally concur in clinical settings, they did, by definition, have to fall within a range of severity that would be amenable to clinical trial inclusion. Thus, cases with obviously invasive cancer on the one hand and a totally normal colon on the other tended to not be included. Consequently, it does not appear that there was any relevant selection bias. Also, these cases did not have a predetermined stage, as assigning a stage was a part of the exercise. As seen in heat maps, the majority of reviewers assigned a particular stage or were within ±1 stage. We calculated interobserver variation and found it to be low. However, we did not assess intraobserver variation. One would expect the intraobserver variability to be smaller in magnitude than the inter-reviewer variability because our current study shows the interreviewer variability is low relative to the intervideo variability, suggesting that even if the intrareviewer variability were measured and assessed, our conclusions on the reliability of the proposed scoring system would not change. However, this would have been an interesting exercise, but submission of replicate videos would have substantially increased the number of videos that would have to be scored, while still providing for the range of cases that were included. The statistical modeling that was performed provided a "best fit" for the approximate number of reviewers and videos that were used and recommended 24 videos and reviewers as the best sample size. Hence, the addition of replicate videos would mean reducing the number of independent videos. In addition, rereview of only "difficult cases" is not advisable because cases that had higher inter-reviewer variability would also be likely to have greater intrareviewer variability, and thus this would give a skewed assessment of the level of intrareviewer variability. Hence, future studies will address this as part of future validation steps. Our study shows that rater-to-rater variability is low, the key initial validation of IPSS. This system can now be implemented to assign polyposis staging in clinical and clinical trial settings. Indeed, the prospective clinical trial setting will be the ideal circumstance in which to more appropriately validate IPSS, also enabling trialists to address some of the unresolved issues here, including the potential importance of (1) intraobserver variation, (2) the amount of time spent reviewing videos, and (3) reconciliation processes for cases in which disagreement exists. Finally, given some of the scatter observed in the more challenging cases, one can only assume that a given reviewer would, on a blinded rereview, upstage or downstage a given case in a fashion similar to that seen in the pool of independent reviewers. This would be an interesting phenomenon to assess and, in fact, will be incorporated into the design of upcoming clinical trials in which IPSS is used.

Depending on how such issues are resolved, it may be feasible to delegate scoring to nonexpert reviewers, after a training period, using the values found here as a benchmark and framework for analysis. Now that we have a framework for analysis, we can also readily see that use of IPSS enables the detection of "outlier" cases in which disagreement in stage assignment exists. These lend themselves well to the development of a reconciliation or adjudication process. Such a process must recognize the absence of perfect fidelity in severity scoring and must provide measures, however arbitrary, for their resolution. In doing so, opportunities to further refine IPSS should emerge.

We did not find differences in scoring by demographic characteristics of the scorers, such as number of patients with FAP per year. We could have collected additional information such as total years in FAP-focused practice, perhaps a better reflection of cumulative FAP experience. However, because most reviewers have been InSiGHT participants for many years, this was not likely an important factor. In addition, there were no real patterns of scoring to indicate that any particular reviewer consistently over- or understaged, relative to the average values provided.

How might the prospective IPSS be applied clinically or in clinical trials, and how would we evaluate its effectiveness? Our initial goal was to develop a system that would establish clinical trial endpoints that closely correspond to the FDA's requirement of clinical benefit. If surgical resection can be delayed because of chemoprevention, with or without polypectomy, then a clinical benefit will have been rendered. Such trials are in preparation, with the IPSS incorporated as a primary endpoint. These trials may also lead to modifications of the IPSS, depending on the results of validation studies using hard clinical outcomes such as surgery, cancer, and death.

Another consequence of the use of the IPSS may be to modify the prevailing standard of care with respect to polyposis intervention; that is, a new staging system may prompt changes in the recommended surveillance intervals or criteria for surgical resection. If one considers the interventions recommended by our panel of experts in the context of actual cases, there is a comfort level with nonsurgical intervention at the earlier stages of adenoma involvement and differing surgical interventions at later stages.

The survey data showed general satisfaction with the IPSS. It suggests that clinicians will develop confidence in the system and that it is easy to use. In addition, the proposed IPSS could affect the criteria for inclusion in trials of chemoprevention in patients with FAP. It is likely that future clinical trial attention will be devoted to patients with intact colons, although historically, we have seen that most chemoprevention trials have limited enrollment to patients who have already undergone colectomy and who have recurrent rectal adenomas.

This is the first step toward developing a staging system for colorectal polyposis. No system is perfect and only improves with time. We believe that this proposal and testing of a colorectal polyposis staging system with stage-specific interventions will enable more reliable measures of patients' response to nonsurgical treatments. In addition, these measures should satisfy the need to determine treatment endpoints that meet the FDA's new requirement of clinical benefit. Further validation of this scoring system can be expected in the course of prospective clinical chemoprevention trials currently in development.

ACKNOWLEDGMENTS

The authors thank all of our reviewers for their time and contribution to this study, SLA Pharma (UK) for their funding support, and the entire team who made this study possible.

REFERENCES

1. Spigelman AD, Williams CB, Talbot IC, et al. Upper gastrointestinal cancer in patients with familial adenomatous polyposis. Lancet 1989;2:783-5.
2. Kennedy RD, Potter DD, Moir CR, et al. The natural history of familial adenomatous polyposis syndrome: a 24 year review of a single center experience in screening, diagnosis, and outcomes. J Pediatr Surg 2014;49:82-6.
3. Levine FR, Coxworth JE, Stevenson DA, et al. Parental attitudes, beliefs, and perceptions about genetic testing for FAP and colorectal cancer surveillance in minors. J Genet Couns 2010;19:269-79.
4. da Luz Moreira A, Church JM, Burke CA. The evolution of prophylactic colorectal surgery for familial adenomatous polyposis. Dis Colon Rectum 2009;52:1481-6.
5. Burt RW, Cannon JA, David DS, et al. Colorectal cancer screening. J Natl Compr Canc Netw 2013;11:1538-75.
6. Grover S, Kastrinos F, Steyerberg EW, et al. Prevalence and phenotypes of APC and MUTYH mutations in patients with multiple colorectal adenomas. JAMA 2012;308:485-92.
7. Steinbach G, Lynch PM, Phillips RK, et al. The effect of celecoxib, a cyclooxygenase-2 inhibitor, in familial adenomatous polyposis. N Engl J Med 2000;342:1946-52.
8. Meyskens FL, Curt GA, Brenner DE, et al. Regulatory approval of cancer risk-reducing (chemopreventive) drugs: moving what we have learned into the clinic. Cancer prevention research 2011;4:311-23.
9. Lynch PM, Burke CA, Phillips R, et al. An international randomised trial of celecoxib versus celecoxib plus difluoromethylornithine in patients with familial adenomatous polyposis. Gut 2016;65:286-95.
10. Cohen J. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull 1968;70:213.
11. Johnson VE, Albert J. Ordinal data modeling (statistics for social science and public policy). New York (NY): Springer-Verlag; 1999.
12. Johnson VE. On Bayesian analysis of multirater ordinal data: an application to automated essay grading. J Am Stat Assoc 1996;91:42-51.
13. Lynch PM, Morris JS, Ross WA, et al. Global quantitative assessment of the colorectal polyp burden in familial adenomatous polyposis by using a web-based tool. Gastrointest Endosc 2013;77:455-63.

(4), Department of Medical and Surgical Sciences, University of Bologna, Bologna, Italy (5), Department of Digestive Tract and General Surgery, Saitama Medical University, Saitama, Japan (6), Division of Surgical Oncology, Hereditary Cancer Registry, Hospital Sirio Libanes, Sao Paulo, Brazil (7), Department of Gastroenterology, Institut de Malalties Digestives i Metabòliques, Hospital Clinic, Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd), Institut d'Investigacions Biomèdiques August Pi I Sunyer (IDIBAPS), University of Barcelona, Barcelona, Spain (8), Division of Internal Medicine, University of Michigan Health System, Ann Arbor, Michigan, USA (9), Section of Colon and Rectal Surgery, Washington University School of Medicine, St. Louis, Missouri, USA (10), Unit of Hereditary Digestive Tract Tumors, Fondazione IRCCS, Istituto Nazionale dei Tumori, Milan, Italy (11); Wolfson Unit for Endoscopy, St. Marks Hospital, Harrow, Middlesex, United Kingdom (12); Department of Internal Medicine, Huntsman Cancer Institute, University of Utah, Salt Lake City, Utah, USA (13), Department of Medical and Surgical Sciences, S. Orsola-Malpighi University Hospital, Bologna, Italy (14), Division of Gastroenterology, Department of Veterans Affairs Eastern Colorado Health Care System and University of Colorado School of Medicine, Denver, Colorado, USA (15), Department of Surgery, Hyogo College of Medicine, Hyogo, Japan (16), The Danish Polyposis Register, Gastrointestinal Unit, Hvidovre University Hospital, Copenhagen, Denmark (17), The Polyposis Registry, St. Mark's Hospital, Harrow, Middlesex, United Kingdom (18), Division of Gastroenterology, Hepatology and Nutrition, Nationwide Children's Hospital, The Ohio State University, Columbus, Ohio, USA (19), Gastroenterology and Nutrition Service, Memorial Sloan-Kettering Cancer Center, New York, New York, USA (20), Department of Surgical Gastroenterology, Aalborg Universitetshospital, Aalborg, Denmark (21), Division of Population Sciences, Division of Gastroenterology, Dana-Farber Cancer Institute, Brigham and Women's Hospital, and Harvard Medical School, Boston, Massachusetts, USA (22), Center for Hereditary Tumors, HELIOS Klinikum Wuppertal, University Witten-Herdecke, Wuppertal, Germany (23).

Reprint requests: Patrick M. Lynch, MD, Department of Gastroenterology, Hepatology, and Nutrition, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Blvd., Houston, TX 77054.

If you would like to chat with an author of this article, you may contact Dr Lynch at plynch@mdanderson.org.

ONLINE APPENDIX

Statistical model

A Bayesian multiple rater model was used to assess concordance of ordinal data across multiple raters. The ordinal score was treated as a latent trait and was modeled with normal distribution, with the latent variables indicating an unmeasured continuous measure of polyposis severity. In particular, define a latent variable α_i that indicates the true polyposis severity score for video i. We assume that rater j's perception of polyp severity is given by t_{ij}, which differs from the true latent polyposis severity score by ε_{ij}. Thus, the rater j perceived latent trait is given by the model $t_{ij} = \alpha_i + \varepsilon_{ij}$, where $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$ represents the rater-to-rater variability. We assume that the α_i are independently distributed normal random variables with variance σ_α^2, $N(0, \sigma_\alpha^2)$, with σ_α^2 indicating the video-to-video variability. Thus, we can define a measure of rater agreement by using $\rho = \sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_\varepsilon^2)$, which is also called the intraclass correlation coefficient (ICC). The measure ρ indicates the proportion of total variability attributed to the video-to-video component and is constrained between 0 and 1. Thus, a higher ρ indicates greater concordance, with $\rho = 1$ indicating that all raters gave the exact same rating to all videos and $\rho = 0.5$ indicating that the variability across raters was equal in magnitude to the variability across videos.

In the latent model, for a score with 5 grades, a total of 4 grade cutoffs must be introduced that link the latent continuous score to the observed ordinal stages. Because the response categories are ordered, we must impose a constraint on the values of grade cutoffs. Given a rater, the ordering constraint may be stated mathematically as $-\infty < \gamma_1 \leq \gamma_2 \leq \gamma_3 \leq \gamma_4 \leq \infty$ (γ_5). When t_{ij} falls between the grade interval $(\gamma_{c\text{-}1}, \gamma_c)$, the observation is classified into category c. The previous distributions were specified as follows: σ_ε^2 has an inverse γ prior, ie, $1/\sigma_\varepsilon^2 \sim \gamma(1, 1)$, and the category cutoffs γ_c are given independent uniform priors. The posterior distributions of $(\sigma_\alpha^2, \sigma_\varepsilon^2, \gamma_c)$ were obtained by using the Markov chain Monte Carlo algorithm. The concordance measure ρ for each posterior sample was calculated from σ_α^2 divided by $\sigma_\alpha^2 + \sigma_\varepsilon^2$, from which the posterior mean and 95% posterior credible intervals were computed. Concordant measures should have $\rho > 0.50$ at a minimum because $\rho \leq 0.5$ suggests that the rater-to-rater variability is of greater magnitude than the video-to-video variability. Thus, as a measure of statistical significance, we computed $P = \text{Prob}\ (\rho \leq 0.50|\text{data})$ as a measure of statistical significance, with $P < .05$ indicating that the level of agreement is significantly greater than this.

Simulation study

We performed a simulation study to determine the necessary sample size to have power to detect a significantly strong concordance of $\rho = 0.70$ and assess the study's operating characteristics under other possible concordances. To simulate data for the studies, we generated multiple datasets based on the different values of σ_α^2 and σ_ε^2 as follows:

- ICC = 0.5 implies $\sigma_\varepsilon^2 = \sigma_\alpha^2$: the rater variation is the same as the video variation
- ICC = 0.67 implies $\sigma_\varepsilon^2 = \sigma_\alpha^2/2$: the rater variation is 1/2 of the video variation
- ICC = 0.75 implies $\sigma_\varepsilon^2 = \sigma_\alpha^2/3$: the rater variation is 1/3 of video variation
- ICC = 0.80 implies $\sigma_\varepsilon^2 = \sigma_\alpha^2/4$: the rater variation is 1/4 of video variation

The procedure of the simulation study can be summarized as follows:

1. Specify J (number of raters) and I (number of videos)
2. Specify a distribution of proportions of each component in the ordered score $p_c (c = 1, 2, 3, 4, 5)$
3. Given a rater and the distribution of pc, obtain 4 cutoffs γ_c from Dirichlet distribution Dir $(5, p_c)$
4. Generate latent trait $t_{ij} = \alpha_i + \varepsilon_{ij}$ as follows:
   i. Generate α_i from normal distribution $N(0, \sigma_\alpha^2)$, where $i = 1,\ldots, I$
   ii. Given α_i, generate ε_{ij} from normal distribution $N(0, \sigma_\varepsilon^2)$, where $j = 1,\ldots, J$
5. Obtain the 5-point ordered score by categorizing t_{ij} by using the cutoffs γ_c^j from Step 3
6. Estimate the posterior distribution of ρ by using the Markov chain Monte Carlo algorithm [1, 2]
7. Claim a significant agreement if Prob $(\rho \leq 0.5|\text{data}) < p_L$, where p_L is the disagreement parameter and should be set low such as 0.05 or 0.1.

Supplemental Table 1 summarizes our simulation result with 5 different scenarios of ICC based on 24 raters and 24 videos from 100 simulated trials, assuming that the distribution of ordered scores p_c are 0.30, 0.25, 0.20, 0.15, and 0.10, respectively. The simulation results showed that a sample size of 24 raters and 24 videos will have at least 83% power for a concordance of an ICC = 0.70 based on this Bayesian multiple-rater modeling. More scenarios with 12 or 18 raters are shown in Supplemental Table 2.

SUPPLEMENTAL TABLE 1. Power estimation based on the number of raters ($J = 24$) and videos ($I = 24$) from 100 simulated trials, assuming that the distribution of ordered scores p_c are 0.30, 0.25, 0.20, 0.15, and 0.10, respectively. We claimed a significant agreement if Prob ($\rho \leq 0.5|\text{data}) < p_L$

Scenarios	$p_L = 0.05$	$p_L = 0.10$	$p_L = 0.20$
$\rho = 0.50$	0.04	0.05	0.07
$\rho = 0.67$	0.69	0.74	0.78
$\rho = 0.70$	0.83	0.88	0.90
$\rho = 0.75$	0.87	0.91	0.93
$\rho = 0.80$	0.99	0.99	1.0

SUPPLEMENTAL TABLE 2. Power estimation based on the number of raters ($J = 12$ or 18) and videos ($I = 20$, 30, or 40) from 100 simulated trials, assuming the distribution of ordered scores p_c are 0.30, 0.25, 0.20, 0.15, and 0.10, respectively. We claimed significant agreement if Prob ($\rho \leq 0.5|\text{data}) < p_L$

Scenario	Power (J = 18/J = 12)		
I = 40	$p_L = 0.05$	$p_L = 0.10$	$p_L = 0.20$
$\rho = 0.50$	0.0/0.0	0.02/0.0	0.02/0.0
$\rho = 0.67$	0.46/0.28	0.54/0.32	0.61/0.41
$\rho = 0.75$	0.87/0.79	0.90/0.79	0.94/0.86
$\rho = 0.80$	0.99/0.95	0.99/0.99	0.99/0.99
I = 30	$p_L = 0.05$	$p_L = 0.10$	$p_L = 0.20$
$\rho = 0.50$	0.0/0.0	0.0/0.0	0.0/0.0
$\rho = 0.67$	0.38/0.35	0.44/0.43	0.52/0.48
$\rho = 0.75$	0.61/0.60	0.69/0.60	0.70/0.70
$\rho = 0.80$	0.82/0.81	0.90/0.84	0.90/0.89
I = 20	$p_L = 0.05$	$p_L = 0.10$	$p_L = 0.20$
$\rho = 0.50$	0.0/0.0	0.0/0.0	0.0/0.0
$\rho = 0.67$	0.14/0.06	0.16/0.07	0.22/0.11
$\rho = 0.75$	0.36/0.31	0.45/0.43	0.50/0.48
$\rho = 0.80$	0.69/0.58	0.71/0.62	0.77/0.67

SUPPLEMENTAL TABLE 3. ICC for IPSS score for video ratings by demographic characteristics

Characteristics	Estimated value (SE)	95% CI
All raters (N = 26)	0.710 (0.027)	0.651-0.759
Specialty		
Surgeon (n = 12)	0.738 (0.039)	(0.654-0.808)
Endoscopist (n = 14)	0.684 (0.037)	(0.604-0.751)
Sex		
Female (n = 6)	0.778 (0.047)	(0.674-0.858)
Male (n = 20)	0.694 (0.032)	(0.631-0.754)
No. of FAP patients		
≤10 (n = 9)	0.743 (0.042)	(0.653; 0.819)
≥11 (n = 17)	0.671 (0.038)	(0.594; 0.741)

ICC, Intraclass correlation; *IPSS*, InSiGHT polyposis staging system; *SE*, standard error; *CI* = confidence interval; *FAP*, familial adenomatous polyposis.

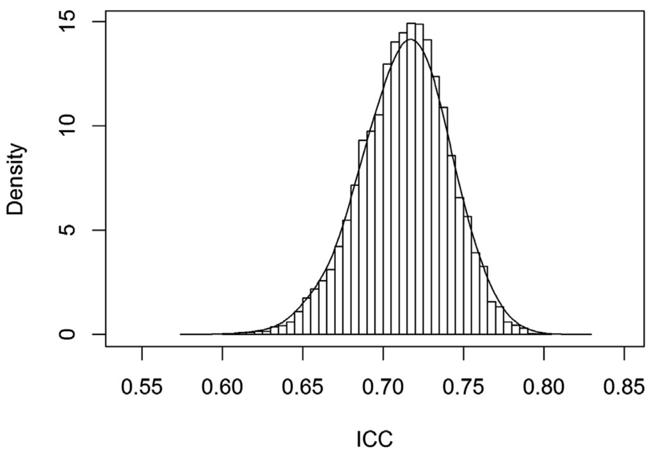
SUPPLEMENTAL TABLE 4. List of questions asked to the reviewers at the end of the reviews (Please provide your opinion on the following statements.)

Surgeon No.	Question	Options	No. (%)
1	The development of a staging system for colorectal polyposis will be helpful in communicating with colleagues regarding patient status.	Strongly agree	18 (69)
		Agree	7 (27)
		Neutral	1 (4)
		Disagree	0
		Strongly disagree	0
2	The development of a staging system for colorectal polyposis will be helpful in evaluating endpoints in clinical chemoprevention trials.	Strongly agree	18 (69)
		Agree	4 (15.5)
		Neutral	4 (15.5)
		Disagree	0
		Strongly disagree	0
3	Subject to my specific comments in the scoring sheet above, I am in general agreement with the present proposed IPSS.	Strongly agree	2 (8)
		Agree	21 (84)
		Neutral	1 (4)
		Disagree	1 (4)
		Strongly disagree	0
4	Subject to my comments on the scoring sheet above, I am in general agreement with the present proposed interventions by stage.	Strongly agree	0
		Agree	16 (62)
		Neutral	8 (30)
		Disagree	2 (8)
		Strongly disagree	0

IPSS, InSiGHT polyposis staging system.

SUPPLEMENTAL TABLE 5. Comment by reviewers on outlier cases

Video	Comments
13	"This one was really difficult, am putting stage 1 because clearly less than 200 polyps. But there are clearly 2 that are >1 cm, so doesn't fit that criteria for stage 1, but does for 2. So it is between stages 1 and 2. Intervention hard too; this is not someone we would consider surgery for. But would do polypectomies of polyps, particularly larger ones at the time of this colonoscopy, and then repeat colonoscopy in 1 year." (Comment 1) "Don't feel Stage 0 is optimal in this case, as feel total polyp count <20, with a single polyp >1 cm. Perhaps offer an alternative stage category for <20 polyps, 1 or more >1 cm, which may better fit this case." (Comment 2) "Tough case. Very few polyps but 1 in ascending colon needs to be removed and would be somewhat dicey endoscopically, especially in pt ultimately destined for colectomy anyway." (Comment 3)
20	"This could be stage 3 too (have difficulty assessing 400 vs 600, etc, by that time it doesn't matter. Saying D because this one is not as severe as some of the others where E is clearly right. But would prefer if this management option was reversed in order. Colectomy or polypectomy of larger polyps and repeat colonoscopy in 6-12 months if desire to avoid surgery." (Comment 1) "Re stage, may consider alternative option of 200-500 polyps, no >1 cm, which may better fit case." (Comment 2)
24	"Is the staging system still applicable to the post-pouch patient? I would biopsy for sure, but there is less certainty that polyps are adenomas (although it certainly is possible). I would feel very uncomfortable making any recommendation regarding further management of the pouch (eg, excision/revision) without histologic information. Suggest that it be pouch polyps and be classified separately." (Comment 1) "I would want to biopsy that area on retroflexion to confirm adenomatous change (does not appear to be overtly adenomatous, but I wouldn't just ignore it), and my follow-up would depend on that path result as well as a few small raised areas elsewhere, although these are likely lymphoid aggregates." (Comment 2)



Supplemental Figure 1. Posterior distribution of the ICC for IPSS score agreement based on 26 raters. The posterior mean (standard error) of the ICC is 0.710 (0.027), with a 95% credible interval between 0.651 and 0.759. *ICC*, intraclass correlation coefficient.