Published in final edited form as:
Gastrointest Endosc. 2013 March ; 77(3): 455–463. doi:10.1016/j.gie.2012.11.038.

Global Quantitative Assessment of Colorectal Polyp Burden in Familial Adenomatous Polyposis Using a Web-based Tool

Patrick M. Lynch[1,*], **Jeffrey S. Morris**[2,*], **William A. Ross**[2], **Miguel A. Rodriguez-Bigas**[3], **Juan Posadas**[6], **Rossa Khalaf**[4], **Diane M. Weber**[5], **Valerie O. Sepeda**[5], **Bernard Levin**[6], and **Imad Shureiqi**[4,5]

[1]Department of Gastroenterology, Hepatology & Nutrition, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA

[2]Department of Biostatistics and Applied Mathematics, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA

[3]Department of Surgical Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA

[4]Department of Gastrointestinal Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA

[5]Department of Clinical Cancer Prevention, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA

[6]Division of Cancer Prevention and Population Sciences, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA

Abstract

Background—Accurate measures of total polyp burden in familial adenomatous polyposis (FAP) are lacking. Current assessment tools include polyp quantitation in limited-field photographs and qualitative total colorectal polyp burden by video.

Objective—To develop global quantitative tools of FAP colorectal adenoma burden.

Design and Interventions—A single-arm phase II trial in 27 FAP patients treated with celecoxib for 6 months, with pre- and post-treatment videos posted to intranet with interactive site for scoring.

Correspondence to: Patrick M. Lynch, JD, MD, Department of Gastroenterology, Hepatology & Nutrition, Unit 1466, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, TX 77030-4009. Phone: (713) 792-5073. Fax: (713) 563-4398. (plynch@mdanderson.org). Imad Shureiqi, MD, Department of Gastrointestinal Medical Oncology, Unit 426, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, TX 77030-4009. Phone: (713) 792-2828. Fax: (713) 745-1163. (ishureiqi@mdanderson.org)..
*Drs. Lynch and Morris contributed equally to this work.

Take-home Message This newly developed web-based polyp scoring method provides a reproducible method to quantify total colorectal adenoma burden in FAP patients. This method improves the endoscopic assessment of this disease to better determine its clinical progression and response to chemopreventive interventions and to develop a clinical staging system for colorectal polyposis.

Authors' contributions: IS, PML, JSM, BL study concept and design; PML, WAR, MAR, JP, DMW, VOS acquisition of data; IS, JSM, PML, WAR, MAR, RK, JP analysis and interpretation of data; IS, JSM, PML drafting of the manuscript; PML, JSM, IS, WAR, MAR, BL critical revision of the manuscript for important intellectual content; JSM, IS statistical analysis; IS, PML, JSM obtained funding; JP web-tool development, implementation and technical support; DMW, VOS. Patient's care support; PML, JSM, IS study supervision.

Potential competing interests: In a separate trial than the one reported in this study, Dr. Lynch received research support from Pfizer and served on the steering committee for that trial. He also served on external advisory panels and speakers bureau for Myriad Genetics. Neither of these activities is directly relevant to this trial.

Main outcome measurements—Global adenoma counts and sizes (grouped into categories: <2 mm, 2–4 mm, and >4 mm) were scored from videos using a novel web-based tool. Baseline and end-of-study adenoma burdens results were summarized using five models. Correlations between pairs of reviewers were analyzed for each model.

RESULTS—Interobserver agreement was high for all 5 measures of polyp burden. Measures employing both polyp count and polyp size had better interobserver agreement than measures based only on polyp count. The measure in which polyp counts were weighted according to diameter, calculated as (1) × (no. of polyps <2 mm) + (3) × (no. of polyps 2–4 mm) + (5) × (no. of polyps >4 mm) had the highest interobserver agreement. (Pearson $r = 0.978$ for two gastroenterologists, 0.786 and 0.846 for the surgeon vs each gastroenterologist). Treatment reduced polyp burden by these measurements in 70–89% subjects ($p<0.001$).

Limitations—Phase II study.

Conclusions—This novel web-based polyp scoring method provides a convenient and reproducible way to quantify global colorectal adenoma burden in FAP patients and a framework for developing a clinical staging system for FAP.

Keywords

colonic polyposis; chemoprevention; gastrointestinal endoscopy; familial adenomatous polyposis

INTRODUCTION and BACKGROUND

Patients with familial adenomatosis polyposis (FAP) develop large numbers of colorectal adenomas.(1) Quantification of polyp burden is important to assess the clinical course of disease and to measure response to chemopreventive interventions.(2, 3) Counting polyps and measuring polyp diameters in real time during colonoscopic procedures is impractical because of the large number of polyps. The methods of quantifying polyp burden commonly used in clinical chemoprevention studies are (1) quantitative measurement of polyp numbers and sizes in still color photographs taken before and after intervention in designated areas of the colorectum and/or (2) qualitative assessment of the total colorectal polyp burden using videotapes of the endoscopic procedure.(2, 3) While the photograph-based approach produces quantitative results and is arguably reproducible, the most important limitation of this approach is that it provides information about only a small field of polyps in the colorectum and does not use all of the information potentially available for assessment of polyp burden. Photograph-based measurement will inherently miss any response outside the chosen field and thus may tend to underestimate response. Video-based measurements capture the entire colon but have been qualitative rather than quantitative—one of a paired set of pre- and postintervention videos is classified as appearing “better than,” “the same as,” or “worse than” its paired video.(2)

Given the limitations of measuring polyp burden with photographs (limited field of view) and videos (findings are qualitative and subjective), we sought to develop a method to quantitate overall polyp burden throughout the colon in patients with FAP using colonoscopy videos covering the entire colorectum. To avoid the difficulty of attempting real-time counting during colonoscopic procedures, we developed a web-based tool to provide scorers the flexibility to assess recorded colonoscopic videos, thus avoiding the problems of attempting real-time counting during colonoscopic procedures. As polyp burden is related to both the number and the size of colorectal polyps, we assessed 5 different methods based on size-specific or “binned” polyp counts to summarize polyp burden. Our goal was to achieve reliable measurements with high correlation between reviewers and low reviewer-to-reviewer variability relative to the subject-to-subject variability. We used this

tool to assess adenoma regression as a measure of the chemopreventive response in a single-arm clinical study of 6-month exposure to celecoxib in patients with FAP.

In designing the trial, we were struck by some of the limitations of the existing methods of scoring polyps, even though we had been involved in the initial development of these methods in the 1990s. Our perception was that improvements in technology might enable development of a more robust method for quantitating polyp burden, not only for use in clinical chemoprevention trials but also to serve as a foundation for assessment of adenoma burden in everyday clinical practice.

METHODS

Study Design and Endoscopic evaluation

Patients with FAP were recruited at The University of Texas MD Anderson Cancer Center. The study was approved by MD Anderson's Institutional Review Board, and patients gave informed consent before participating. This single-arm celecoxib study enrolled 47 patients between November 2004 and May 2010. A baseline colonoscopy (or sigmoidoscopy in patients who had undergone colectomy) was performed before initiation of celecoxib, and a follow-up colonoscopy/sigmoidoscopy was performed after 6 months of celecoxib treatment. The celecoxib dose was 400 mg by mouth twice daily for 6 months. The eligibility criteria are as described in ClinicalTrials.gov number NCT00503035 and as a part of the detailed study clinical design and endoscopic evaluation sections in Supplemental Methods.

Development of Web-based Scoring Tool

We developed a secure web-based tool with the necessary elements for efficient, unbiased review of videos from colonoscopy procedures. These web program elements consisted of: the list of de-identified videos, the colonoscopy video (streaming), and the scoring field (Figure 1). We implemented a web-based solution that uses a browser supporting Windows Media Player to deliver the high-definition videos. We utilized Microsoft Technologies to develop the Web-user interface (using ASP.Net) and the storing engine (MS SQL Server). Colonoscopy videos were captured in DVD format, edited, provided in high-definition windows media file format (.wmv), and stored on a file server. Information related to each video, such as file location and date and time recorded, was stored in a database where it related to other study-specific datasets. Each investigator, upon accessing the web program, was provided with the list of videos to be scored, with the videos coded in order to blind the reviewer to specific patient information or treatment state (pre- vs post-celecoxib). The coded videos were reviewed in a random order; thus, the precelecoxib videos were not all scored before the post-celecoxib videos. Each reviewer completed a scoring field for each video.

Polyp Burden Scoring

In the interest of efficiency in scoring, all videos were edited by one of the reviewers (PML) before scoring to delete extraneous material (e.g., an initial partial pull-back that was aborted and restarted but captured on video; aspiration of retained prep material). Care was taken to capture all polyps in a segment of colon despite any necessary editing. Following such editing, videos were posted to a shared drive and then loaded to the web-based scoring tool. For each video segment (cecum-ascending, transverse, descending-sigmoid, and rectum), polyps were counted within each of 3 diameter categories (see video about use of web tool in supplementary information): <2 mm, 2–4 mm, and >4 mm. By assembling counts for different diameter groups, we were able to construct measures of polyp burden that take polyp size into account. By limiting the number of diameter groups to 3, we were

able to keep the counting process efficient and at a scale that can be estimated by the eye without the need for formal measurement. As the reviewer viewed a video, the mouse was clicked in the appropriately binned polyp-diameter field (<2mm, 2–4mm, >4mm) to mark each polyp (Figure 1). Each mouse click increased the polyp count in the field by 1, similar to use of an abacus or manual cell counter. Importantly, the video could be paused as needed to count a large number of polyps in a given field and then restarted for continued review. Each video was scored by each of 3 study investigators (R1, R2, R3): two of these investigators, a gastroenterologist (R2) and a surgeon (R1), are highly experienced in the management of patients with FAP. R3 is an experienced general gastroenterologist who had performed some of the study procedures, while R2 performed the bulk of them.

Polyp Burden Measures

We considered 5 different methods for summarizing a patient's total polyp burden from the video-based counts—2 measures based on polyp count only and 3 measures based on both polyp count and polyp size. The first two (TC, TC2) simply sum up polyp counts, with TC summing over all counts, and TC2 only summing over counts of medium (2–4mm) and large (>4mm) polyps. The other three methods (ES, MR, BR) take weighted averages, with the counts in larger polyp classes counting more. These three methods differ in terms of the relative weights given to small, medium, and large polyps in computing the total polyp burden. Let C_S, C_M, and C_L be the counts of small (<2 mm), medium (2–4 mm), and large (>4 mm) polyps, respectively. The “equally spaced” weighted-counts measure (ESWC, or “ES” for brevity) used equally spaced weights across the 3 categories: with ES = $(1\times C_S) + (2\times C_M) + (3\times C_L)$. The “mid-range” weighted-counts measure (MR) roughly used the categories' mid-range value for the weights: with MR = $(1\times C_S) + (3\times C_M) + (5\times C_L)$. The “bottom-range” weighted-counts measure (BR) used the categories' bottom-range values for the weight: with BR = $(1\times C_S) + (2\times C_M) + (4\times C_L)$. We computed each of these 5 measures from the counts recorded by each reviewer for each video, which corresponded to 1 region from 1 patient at 1 time point (either baseline or posttreatment). The statistical assessment of polyp burden measures is as described in Supplemental Methods.

RESULTS

Patients' Demographic and Clinical Characteristics

The clinical study flow is summarized in Figure 2. In all, 47 patients were enrolled in the study. Of these, 15 were disqualified from continuing on the study: 2 had uncontrolled hypertension, 4 had uncontrolled hyperlipidemia, 1 had leukopenia detected on study-screening laboratory tests, 7 had insufficient polyp numbers for tissue sampling as specified by the protocol; and 1 had colon cancer detected on the initial colonoscopy. Three patients did not complete the study: 1 had FAP-related symptoms of abdominal pain and rectal bleeding necessitating total proctocolectomy, and 2 were lost to follow-up after the baseline colonoscopy. Thus, 29 patients completed the 6 months of celecoxib treatment. Of these 29 patients, 27 had fully comparable pairs of pre- and posttreatment videos. Of the 27 fully evaluable patients, 22 had an intact colorectum, while 5 had undergone previous colectomy with ileorectal anastomosis.

Patients' demographics are shown in Supplementary Table 1. The original-cohort gender distribution was 50% for each gender. For the patients who completed the study, the gender distribution was 40% females and 60% males. Ethnicity distribution for the original cohort was African-American 1 (2%), Hispanic 9 (19%), and white 38 (79%).

Adherence to Treatment and Side Effects of Treatment

The average rate of adherence for taking celecoxib doses was 94.51% (SD: 5.5%); Celecoxib was well tolerated. Further details for treatment adherence and side effects are described in the supplementary result section.

Inter-reviewer Reliability

Table 1 summarizes the inter-reviewer reliability in terms of Pearson correlation coefficients for each pair of reviewers; corresponding Spearman rank correlation coefficients are provided in Supplementary Table 2. We observed high correlations (0.803–0.978) between the reviewers' estimates of the various polyp burden measures, indicating strong inter-reviewer reliability. These correlations were especially strong between the 2 gastroenterologists (correlations 0.957–0.978), while the correlations between the surgical reviewer and each gastroenterologist were not quite as high (0.803–0.882). Similar results were obtained for Spearman correlations (Supplementary Table 2). Generally, the polyp burden measures based on both polyp size and polyp count (ES, MR, and BR) had higher correlations (0.844–0.978) than those based on polyp count only (TC, TC2, 0.803–0.969). The various measures (ES, MR, BR) based on both polyp size and polyp count were nearly equivalent, although both Pearson and Spearman correlations were slightly higher for the MR (mid-range weighted-counts) measure.

On analysis of correlation between reviewers by region, we found that between-reviewer correlations seemed to be slightly higher in the rectal and descending colon region than in the other regions and slightly lower in the ascending colon region than in the other regions (Supplementary Tables 3–10).

Table 2 presents the results from the linear mixed model analyses for each polyp burden measure. The table includes restricted maximum likelihood estimates of the variance components. Recall that the subject-to-subject variability represents the variance across the videos from different patients; the reviewer-to-reviewer variance represents the variance across systematic effects for each reviewer averaged over videos/patients, which could be called *systematic reviewer variability*; and the residual error represents reviewer-to-reviewer variability as it differs across videos/patients, which could be called *non-systematic reviewer variability*.

We see that for all measures, the subject-to-subject variability (72–82% of total variability) was much greater than the systematic reviewer variability (1–5% of total variability) or non-systematic reviewer variability (residual error, 17–25% of variability). Summing together the systematic and non-systematic reviewer variabilities, we see that 18–28% of variability in the measurements was due to reviewer-related effects, while 72–82% of the variability in the data was from natural subject-to-subject (biological) variability. Again, we see that the polyp burden measures based on both polyp size and polyp count showed greater inter-reviewer reliability than those involving only polyp count. These results suggested that these polyp burden measures had strong inter-reviewer reliability.

A later modification of the web-based scoring tool allowed us to record the length of the edited videos as well as the time taken by each reviewer to score polyps, including times the video was paused or rewound. These times were summed across regions within patients to obtain a video length for each patient. These video scoring duration data were available for both pre-celecoxib and post-celecoxib endoscopies for 81% of patients and for either pre-celecoxib or post-celecoxib endoscopy but not both for another 15%. Statistical summaries of the video lengths are provided in Table 3, along with *normalized review time* for each reviewer, defined as the review time divided by the edited video length. From this analysis, we see that the mean video length was 212.5 seconds. Across all reviewers, the median

normalized review time was 3.2, meaning that the reviewer took 3.2 times the length of the video to do the reviewing, when pause and rewind time was included. The normalized review times ranged from a minimum of 1.1 to a maximum of 19.1. We found that the gastroenterologists tended to take considerably more time to do the reviews than did the surgeon (median normalized review times of 3.0 and 3.8 for the 2 gastroenterologists and 2.1 for the surgeon).

Effects of Celecoxib on Polyp Burden

After 6 months of celecoxib treatment, mean polyp burden decreased between 13.27% and 39.46%, depending on the polyp burden assessment measure and the reviewer (Table 4). The polyp burden was judged to have decreased compared with baseline in the vast majority of the patients regardless of the measure or reviewer (Figure 3, Supplementary Table 11, $p<0.001$) The percentages of patients with 25% or more reduction in their polyp burden (Supplementary Table 12) ranged from 53% to 85% across the various polyp burden measures and reviewers.

DISCUSSION

Findings from the current study demonstrate that a web-based scoring tool can be used to quantify polyp burden in colonoscopy or sigmoidoscopy videos from patients with FAP in a straightforward and reproducible fashion. Videos were relatively short (approximate median time of 3.4 minutes), and review time was manageable (approximate median of 11 minutes per patient). For all 5 measures of total polyp burden calculated on the basis of the scoring data recorded by the reviewers.—2 measures based on polyp count only and 3 measures based on both polyp count and polyp size—we observed high inter-reviewer reliability.

We consistently found that the measures based on both polyp count and polyp size yielded more reliable measurements. Since these measures capture more information than the measures based on polyp count alone, we expect them to also have greater validity and thus greater precision as a measure of total polyp burden. This suggests that there are benefits of our strategy of having the reviewers' score on the basis of both polyp size and polyp number and not just on the basis of polyp number. This study cannot tell us whether the 3 size categories that we used are better or worse than alternatives that might have been employed —for example, 2 size categories or 4 size categories with different cut-offs for polyp diameter. Our choice of 3 size categories was guided by our desire to have reviewers measure the polyp size as well as possible while not burdening the reviewers too much and not asking them to distinguish between 2 size categories that cannot be easily distinguished by the naked eye.

From our variance components analysis, we found that the vast majority of the variability in the data (75–80% or so) was subject-to-subject variability; only approximately 20% of the variability came from reviewer-related sources. Thus, the ratio of “biological” to “technical” variability was roughly 8/2=4, further suggesting that our web-based scoring tool leads to reasonably reliable measurements of polyp burden. Of the 5 measures of polyp burden, the measure with the highest relative technical variability was the one that defined total polyp burden as the sum of the counts of all tumors ≥2 mm; the other 4 measures had similar, considerably lower technical variability.

While the inter-reviewer reliability was high, there was still evidence of reviewer-to-reviewer variability, with systematic reviewer effects accounting for approximately 1–5% of the variability and non-systematic reviewer effects accounting for another 15–20% of the variability. The existence of these reviewer effects indicates that it is essential when scoring to attempt to control the reviewer-related variability. For example, if the key comparison is

between baseline and posttreatment, as in this study, the same reviewer should rate both baseline and posttreatment polyp burden for a given patient. Since the counts are based on the digitized videos taken during colonoscopy, it is similarly important to have the same endoscopist perform the baseline and posttreatment colonoscopies to maximize similarity in the colonoscopy videos.

If the systematic and non-systematic reviewer variabilities had both been 0, indicating that all reviewers obtained exactly the same measurements from the same video and reviewers repeatedly looking at the same video achieved the same measurements, then it would be sufficient to have just 1 reviewer score each video once. However, because there is at least *some* reviewer-to-reviewer variability, it is advisable to have multiple reviewers independently score each video and then average the individual scores to reduce the total variability in the polyp burden estimate. This averaging would lead to reduced measurement error in polyp burden estimates and thus more power for detecting differences between treatment groups in terms of average polyp burden.

The current study is the first, to our knowledge, to utilize a quantitative assessment of the polyp burden in the entire colorectum. Previous studies have relied on quantitative assessment of selected photographs, qualitative video-based assessment of the entire colorectum, (2, 3) or real-time counting performed by the endoscopist. The photographs provide quantitative measurements but for only a small part of the colon, and may vary depending of depth of field unless a measuring tool (open or closed forceps) is placed immediately adjacent to each measured polyp. Global assessments may cover the entire colon but are generally qualitative. In our original trial evaluating celecoxib in patients with FAP, [2] reviewers blinded as to whether videos were from before or after treatment simply rated one video as better as or worse than its paired mate. At that time, we had no convenient way to capture and review digitized videos. The video-based methods employed in the current study reflect the ability to routinely capture, edit, maintain, and post for convenient web-based review large volumes of information. In the case of colonoscopy video, the approach employed is both global and quantitative, so it has the potential to more accurately capture a patient's total polyp burden. In our original celecoxib trial[2] video reviewers met together for several days to review a volume of videos, involving considerable expense as well as fatigue. In the current trial, scorers were able to review videos at their own pace in the comfort of their office or at home, with secure transmission of scores for polyp burden.

For 2 of the 3 reviewers in this study, the estimated mean reduction in total polyp burden was higher (36.05–39.36%) than the 28% reduction in polyp numbers in our previous trial of celecoxib in which response was evaluated using selected photographs.[(2)] However, treatment schedule, procedure, and patient population were very similar between the current study and our prior study.(2) Interestingly, the reduction in total polyp number, regardless of size (28.59–30.67%), was similar to the figure from the prior study (28%). These findings could be interpreted to suggest that assessment of polyp numbers without accounting for size has the potential to underestimate the effect of celecoxib on polyp burden.

The polyp burden measurements provided by the surgeon in the current study were significantly lower than those of the 2 gastroenterologists and lower than the measurements in our prior celecoxib study,(2) raising the possibility that the assessment of response can be affected by the speciality of the reviewer. Also, the surgeon devoted less time, on average, per video, than did the gastroenterologists. Thus, the difference in polyp counts could be due to differences in review duration. This could be more formally evaluated in future studies with larger numbers of reviewers of different specialties.

In our prior study of celecoxib, 53% of patients who received 400 mg of celecoxib twice daily had a 25% or greater reduction in the mean number of polyps.(2) In the current study, the proportions of patients with a 25% or greater reduction in the mean number of polyps were as follows: gastroenterologists, 63–85%; surgeon, 53–63%. These findings suggest that the proportion of patients with FAP who benefit from celecoxib might be higher than previously reported.

While FAP remains a rare disease, managing these patients poses challenges that practitioners could face anywhere. As management in centers of excellence evolves toward more commonly postponing colectomy or proctocolectomies, use of existing and emerging chemopreventive agents can be expected to increase and result in more challenging endoscopic surveillance. Given the great mobility of the population and frequent moves of individuals between states if not countries, the availability of a tool to accurately quantify and communicate endoscopic findings among endoscopists is highly desirable.

This new web-based scoring tool provides a convenient method for reproducible, longitudinal assessments of changes in the polyp burden in the entire colorectum in patients with FAP. This new web tool could allow gastroenterologists in academic and non-academic settings to not only post videos of colonoscopies but assess the findings using a rationally developed quantitative method to better communicate endoscopic results for FAP patients.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Financial Support: This work was supported in part by National Cancer Institute grant R01s CA106577 and CA137213, the Caroline Wiess Law Endowment for Cancer Prevention, a National Colorectal Cancer Research Alliance grant, and the National Institutes of Health through MD Anderson's Cancer Center Support Grant, CA016672. The funding agencies had no involvement in the design of the study; the collection, analysis, or interpretation of the data; the writing of the manuscript, and the decision to submit the manuscript for publication. The authors were scientifically fully independent in the conduct of the study and these activities.

REFERENCES

1. Lynch P. Standards of care in diagnosis and testing for hereditary colon cancer. Familial Cancer. 2008; 7(1):65–72. [PubMed: 17701450]
2. Steinbach G, Lynch PM, Phillips RK, Wallace MH, Hawk E, Gordon GB, et al. The effect of celecoxib, a cyclooxygenase-2 inhibitor, in familial adenomatous polyposis. N Engl J Med. 2000; 342(26):1946–52. [PubMed: 10874062]
3. West NJ, Clark SK, Phillips RKS, Hutchinson JM, Leicester RJ, Belluzzi A, et al. Eicosapentaenoic acid reduces rectal polyp number and size in familial adenomatous polyposis. Gut. Jul 1; 2010 59(7):918–25. 2010. [PubMed: 20348368]

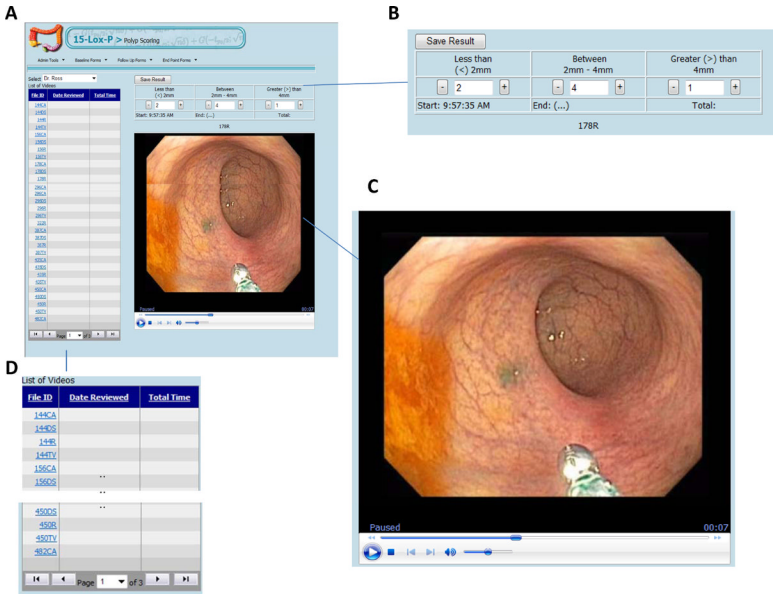

Figure 1.

Web-based tool for scoring polyp burden. A. Picture of the scoring screen for the web-based tool. B–D. Higher-magnification views of the components of the scoring screen: the scoring card (B), streaming video (C), and list of videos (D).

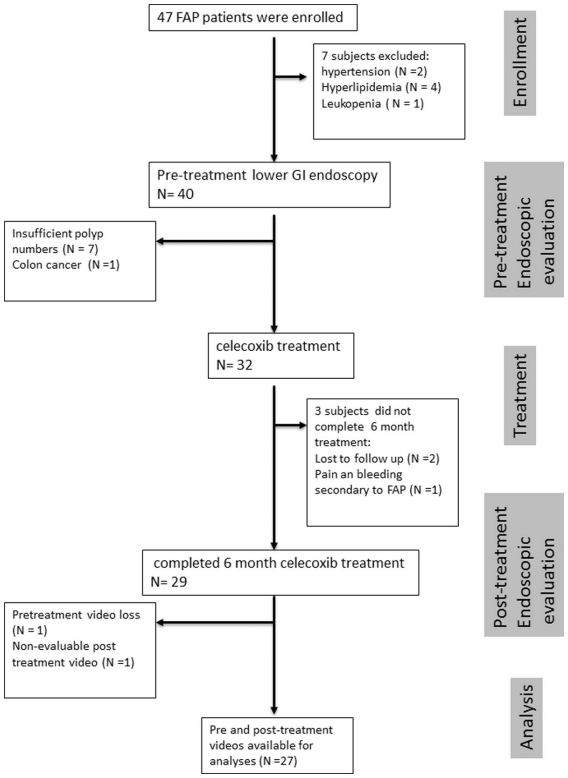

Figure 2.
Study flow diagram. FAP, familial adenomatosis polyposis.

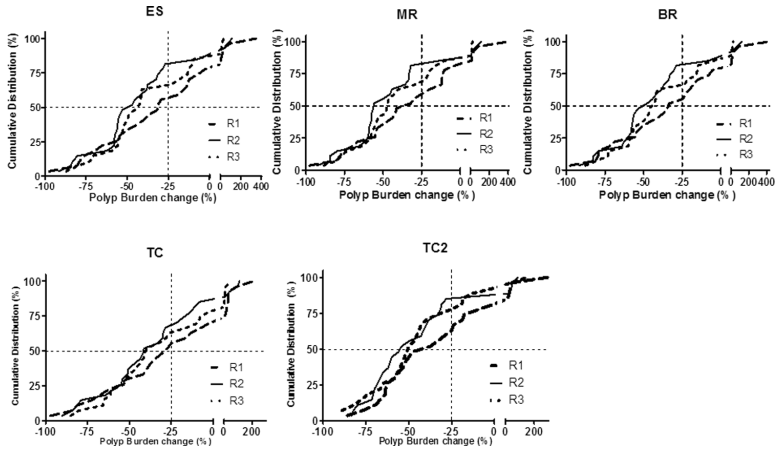

Figure 3.

Cumulative distribution of per cent changes in total polyp burden following 6 months of celecoxib treatment by reviewer and polyp burden measure. In the top 3 panels, total polyp burden was calculated as follows, where C_S, C_M, and C_L were the counts of small (<2 mm), medium (2–4 mm), and large (>4 mm) polyps, respectively: equally spaced "ES", polyp burden = $(1\times C_S)+(2\times C_M)+(3\times C_L)$; mid-range "MR", polyp burden = $(1\times C_S)+(3\times C_M)+(5\times C_L)$; and bottom-range "BR", polyp burden = $(1\times C_S)+(2\times C_M)+(4\times C_L)$. TC "Total Count" indicates that total polyp burden was determined by summing the counts for polyps in all 3 size groups; TC2 "Count 2" indicates that total polyp burden was determined by summing the counts for medium (2–4 mm) and large (>4 mm) polyps only.

Table 1

Inter-reviewer reliability *

Polyp Burden Measure †	R1 ‡	R2	R3	
ES	1.000	**0.848**	**0.876**	**R1**
		1.000	**0.977**	**R2**
			1.000	**R3**
MR	1.000	**0.846**	**0.867**	**R1**
		1.000	**0.978**	**R2**
			1.000	**R3**
BR	1.000	**0.844**	**0.874**	**R1**
		1.000	**0.977**	**R2**
			1.000	**R3**
TC	1.000	**0.841**	**0.882**	**R1**
		1.000	**0.969**	**R2**
			1.000	**R3**
TC2	1.000	**0.837**	**0.803**	**R1**
		1.000	**0.957**	**R2**
			1.000	**R3**

* Pearson correlations between reviewers were calculated with respect to the square root transform of total polyp burden estimates, combined over regions within each patient and time period.

† See Methods section for formulas for calculating the following polyp burden measures: equally spaced weighted counts (ES), mid-range weighted counts (MR), bottom-range weighted counts (BR). TC = sum of polyp counts in all size categories; TC2 = sum of polyp counts in the 2–4 mm and >4 mm categories.

‡ R: reviewer.

Table 2

Variance components for video scoring*

Polyp Burden Measure†	Subject-to-Subject variance (relative variability)	Reviewer-to-Reviewer variance (relative variability)	Residual variance (relative variability)
ES	797 (0.82)	14 (0.01)	162 (0.17)
MR	1689 (0.81)	22 (0.01)	387 (0.18)
BR	975 (0.81)	18 (0.01)	215 (0.18)
TC	243 (0.78)	14 (0.05)	53 (0.17)
TC2	97 (0.72)	3 (0.02)	34 (0.25)

The model included variance components for subject-to-subject variance (biological variability), reviewer-to-reviewer variance (*systematic reviewer variability*) and residual variance (*non-systematic reviewer variability*). The table presents restricted maximum likelihood (REML) estimates of the variance components along with relative proportion of variability at that level in parenthesis. High interrater reliability is indicated by high proportions of variability coming from the natural subject-to-subject variability relative to the technical reviewer-related sources.

*Variance components from linear mixed model analysis done on the various polyp burden measures.

†See Methods section for the formulas for calculating the following polyp burden measures: equally spaced weighted counts (ES), mid-range weighted counts (MR), and bottom-range weighted counts (BR). TC = sum of polyp counts in all size categories; TC2 = sum of polyp counts in the 2–4 mm and >4 mm categories.

Table 3

Video scoring duration by reviewer

		Min	Q1‡	Median	Q3‡	Max	Mean	SD
Video Length (*sec*) *		73.0	137.3	203.5	277.3	384.0	212.5	89.2
Normalized Review Time †	***R1***	1.1	1.6	2.1	3.7	7.4	2.8	1.5
	R2	1.1	2.6	3.0	4.4	19.1	4.3	3.6
	R3	2.1	3.1	3.8	4.4	14.1	4.5	2.6
	All Reviewers	1.1	2.3	3.2	4.2	19.1	3.9	2.8

*The video length is computed for each subject, summing the total length of the edited video across all available regions.

†The video review times are normalized, meaning they are computed as ratios of the actual total time of the reviewer (R) to score the video divided by the total review time. Thus a video review time of 2.0 means the time they actually spent scoring the video, including any pauses or rewinds, was double the actual length of the video if played without stopping.

‡For each measure, we summarize by the minimum (min), maximum (max), mean, median, first quartile (Q1), third quartile (Q3), and standard deviation (SD).

Table 4

Percentage change in polyp burden by reviewer and measure

Polyp Burden Measure*	R1†	R2	R3
ES	−15.09 (−47.94 to 17.75)‡	−36.05 (−54.08 to −18.02)	−36.19 (−48.12 to −24.27)
MR	−15.05 (−51.60 to 21.49)	−38.66 (−56.62 to −20.70)	−38.94 (−50.9 to −26.9813)
BR	−13.27 (−51.73 to 25.18)	−36.99 (−54.99 to −18.98)	−36.92 (−48.93 to −24.91)
TC	−14.8604 (−39.87 to 10.15)	−28.59 (−47.78 to −9.40)	−30.67 (−43.50 to −17.84)
TC2	−21.08 (−49.51 to 7.36)	−39.46 (−56.69 to −22.24)	−39.22 (−57.83 to −20.60)

*See Methods section for formulas for calculating the following polyp burden measures: equally spaced weighted counts (ES), mid-range weighted counts (MR), and bottom-range weighted counts (BR). TC = sum of polyp counts in all size categories; TC2 = sum of polyp counts in the 2–4 mm and >4 mm categories.

†R: reviewer

‡Values are means (95% confidence interval).