January 2013

Global quantitative assessment of the colorectal polyp burden in familial adenomatous polyposis using a Web-based tool
Global quantitative assessment of the colorectal polyp burden in familial adenomatous polyposis by using a Web-based tool

Patrick M. Lynch, JD, MD,*,1 Jeffrey S. Morris, PhD,*,2 William A. Ross, MD, MBA,2 Miguel A. Rodriguez-Bigas, MD,3 Juan Posadas, MBA, BS,6 Rossa Khalaf, MD,4 Diane M. Weber, RN, BSN,5 Valerie O. Sepeda, RN, BSN,5 Bernard Levin, MD,6 Imad Shureiqi, MD, MS4,5
Houston, Texas, USA

Background: Accurate measures of the 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 the FAP colorectal adenoma burden.

Design: A single-arm, phase II trial.

Patients: Twenty-seven patients with FAP.

Intervention: Treatment with celecoxib for 6 months, with before-treatment and after-treatment videos posted to an intranet with an interactive site for scoring.

Main Outcome Measurements: Global adenoma counts and sizes (grouped into categories: <2 mm, 2-4 mm, and >4 mm) were scored from videos by using a novel Web-based tool. Baseline and end-of-study adenoma burden results were summarized by using 5 models. Correlations between pairs of reviewers were analyzed for each model.

Results: Interobserver agreement was high for all 5 measures of polyp burden. Measures that used 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 $(1) \times \text{(no. of polyps <2 mm)} + (3) \times \text{(no. of polyps 2-4 mm)} + (5) \times \text{(no. of polyps >4 mm)}$ 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 the polyp burden by these measurements in 70% to 89% of patients ($P < .001$).

Limitations: Phase II study.

Conclusion: This novel, Web-based polyp scoring method provides a convenient and reproducible way to quantify the global colorectal adenoma burden in FAP patients and a framework for developing a clinical staging system for FAP.

(Gastrointest Endosc 2013;xx:xxx.)

Abbreviations: BR, bottom-range weighted–counts measure; CL, counts of large (>$4$ mm) polyps; CM, counts of medium (2-4 mm) polyps; CS, counts of small (<$2$ mm) polyps; ES, equally spaced weighted–counts measure (ESWC, or ES for brevity); FAP, familial adenomatous polyposis; MR, mid-range weighted–counts measure; R, reviewer; TC, total count, sum of polyp counts in all size categories; TC2, second count, sum of polyp counts in the 2-4-mm and >4-mm categories.

DISCLOSURE: This work was supported in part by National Cancer Institute grants 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 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, or the decision to submit the manuscript for publication. The authors were scientifically fully independent in the conduct of the study and these activities. P. Lynch is a steering committee member for Pfizer in unrelated clinical studies and served on external advisory panels and speakers bureau for Myriad Genetics. Neither of these activities is directly relevant to this trial. No other financial relationships relevant to this publication were disclosed.

*Drs Lynch and Morris contributed equally to this article.

Copyright © 2013 by the American Society for Gastrointestinal Endoscopy 0016-5107/$36.00 http://dx.doi.org/10.1016/j.gie.2012.11.038

Received September 6, 2012. Accepted November 27, 2012.

Current affiliations: Department of Gastroenterology, Hepatology & Nutrition (1), Department of Biostatistics and Applied Mathematics (2), Department of Surgical Oncology (3), Department of Gastrointestinal Medical Oncology (4), Department of Clinical Cancer Prevention (5), Division of Cancer Prevention and Population Sciences (6), The University of Texas MD Anderson Cancer Center, Houston, Texas USA.

Reprint requests: 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.

If you would like to chat with an author of this article, you may contact Dr Lynch at plynch@mdanderson.org.
Patients with familial adenomatous polyposis (FAP) develop large numbers of colorectal adenomas. Quantification of the polyp burden is important for assessing the clinical course of disease and measuring responses to chemopreventive interventions. Although 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 the 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 before-intervention and after-intervention videos is classified as appearing “better than,” “the same as,” or “worse than” its paired video.

Given the limitations of measuring the polyp burden with photographs (limited field of view) and videos (findings are qualitative and subjective), we sought to develop a method to quantify the overall polyp burden throughout the colon in patients with FAP by 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. Because the polyp burden is related to both the number and size of colorectal polyps, we assessed 5 different methods based on size-specific or “binned” polyp counts to summarize the polyp burden. Our goal was to achieve reliable measurements with high correlation between reviewers and low reviewer-to-reviewer variability relative to the patient-to-patient 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 the polyp burden, not only for use in clinical chemoprevention trials but also to serve as a foundation for assessment of the adenoma burden in everyday clinical practice.

Take-home Message
- This newly developed, Web-based polyp scoring method provides a reproducible method to quantify the total colorectal adenoma burden in familial adenomatous polyposis patients.
- This method improves the endoscopic assessment of this disease to better determine its clinical progression and response to chemopreventive interventions. It also provides a foundation for developing a clinical staging system for colorectal polyposis.

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 and/or 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 (available online at www.giejournal.org).

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 (Fig. 1). We implemented a Web-based solution that uses a browser supporting Windows Media Player to deliver the high-definition videos. We used Microsoft Technologies to develop the Web-user interface (using ASP.Net) and the storing engine (MS SQL Server). Colonoscopy videos were captured in digital video disk 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, on 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 (before vs after celecoxib). The coded videos were reviewed in random order; thus, the before-celecoxib videos were not all
scored before the after-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 (P.M.L.) before scoring to delete extraneous material (eg, an initial partial pull-back that was aborted and restarted but captured on video; aspiration of retained preparation material). Care was taken to capture all polyps in a segment of colon despite any necessary editing. After 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; rectum), polyps were counted within each of 3 diameter categories (See supplementary material available online at www.giejournal.org): <2 mm, 2-4 mm, >4 mm. By assembling counts for different diameter groups, we were able to construct measures of the 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 (<2 mm, 2-4 mm, >4 mm) to mark each polyp (Fig. 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 be restarted for continued review. Each video was scored by each of 3 study investigators (reviewer 1 [R1], reviewer 2 [R2], reviewer 3 [R3]): two of these investigators, a gastroenterologist (R2) and a surgeon (R1), are highly experienced in the management of patients with FAP. Reviewer 3 is an experienced general gastroenterologist who had performed some of the study procedures, whereas 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 the polyp count only and 3 measures based on both the polyp count and polyp size. The first 2 (total count, sum of polyp counts in all size
categories [TC] and count 2, sum of polyp counts in the 2-4-mm and >4-mm categories [TC2]) simply sum up polyp counts, with TC summing over all counts and TC2 summing only over counts of medium (2-4 mm) and large (>4 mm) polyps. The other 3 methods (equally spaced weighted–counts measure [ES], mid-range weighted–counts measure [MR], and bottom range weighted–counts measure [BR]) take weighted averages, with the counts in larger polyp classes counting more. These 3 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 (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 ES weighted–counts measure used equally spaced weights across the 3 categories with \( ES = (1 \times C_S) + (2 \times C_M) + (3 \times C_L) \). The MR weighted–counts measure roughly used the categories’ mid-range values for the weights, with \( MR = (1 \times C_S) + (3 \times C_M) + (5 \times C_L) \). The BR weighted–counts measure 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 the polyp burden measures is as described in Supplemental Methods (Available online at www.giejournal.org).

RESULTS

Patient 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 in 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 before-treatment and after-treatment videos. Of the 27 fully evaluable patients, 22 had an intact colorectum, whereas 5 had undergone previous colectomy with ileorectal anastomosis.

Patient demographic characteristics are shown in Supplementary Table 1 (available online at www.giejournal.org). The original-cohort sex distribution was 50% for each sex. For the patients who completed the study, the sex distribution was 40% female and 60% male.

Adherence to treatment and side effects of treatment

The average rate of adherence for taking celecoxib doses was 94.51% (standard deviation 5.5%); celecoxib was well-tolerated. Further details for treatment adherence and side effects are described in the supplementary result section (See supplementary material available online at www.giejournal.org).

Interobserver reliability

Table 1 summarizes the interobserver reliability in terms of Pearson correlation coefficients for each pair of reviewers; corresponding Spearman rank correlation coefficients are provided in Supplementary Table 2 (See supplementary material available online at www.giejournal.org). 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), whereas 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, available online at www.giejournal.org). 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 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, available online at www.giejournal.org).

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 patient-to-patient 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 and/or patients, which could be called systematic reviewer variability, and the residual error represents reviewer-to-reviewer variability as it differs across videos and/or patients, which could be called nonsystematic reviewer variability.

We see that for all measures, the patient-to-patient variability (72%-82% of total variability) was much greater than the systematic reviewer variability (1%-5% of total variability) or nonsystematic reviewer variability (residual error, 17%-25% of variability). Summing together the systematic
and nonsystematic reviewer variabilities, we see that 18% to 28% of variability in the measurements was due to reviewer-related effects, whereas 72% to 82% of the variability in the data was from natural patient-to-patient (biological) variability. Again, we see that the polyp burden measures based on both polyp size and polyp count showed greater interobserver reliability than those involving only polyp count. These results suggested that these polyp burden measures had strong interobserver 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 before-celecoxib and after-celecoxib endoscopies for 81% of patients and for either before-celecoxib or after-celecoxib endoscopy but not both for another 15% of patients. Statistical sum-
Effects of celecoxib on the polyp burden

After 6 months of celecoxib treatment, the 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 the baseline in the vast majority of the patients, regardless of the measure or reviewer (Fig. 3, Supplementary Table 11; P < .001; available online at www.giejournal.org). The percentages of patients with 25% or more reduction in the polyp burden (Supplementary Table 12, available online at www.giejournal.org) 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 the 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 interobserver reliability.

We consistently found that the measures based on both polyp count and polyp size yielded more reliable measurements. Because these measures capture more information than the measures based on polyp count alone, we expect them to have greater validity and thus greater precision as a measure of total polyp burden. This suggests that there are benefits to 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 used—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 distinguished easily 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 patient-to-patient 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 the polyp burden. Of the 5 measures of the 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.

Although the interobserver reliability was high, there was still evidence of reviewer-to-reviewer variability, with systematic reviewer effects accounting for approximately 1% to 5% of the variability and nonsystematic reviewer

Table 1. Interobserver reliability

<table>
<thead>
<tr>
<th>Polyp burden measure</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ES</td>
<td>1.00</td>
<td>0.848</td>
<td>0.876</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>0.977</td>
<td></td>
</tr>
<tr>
<td>MR</td>
<td>1.00</td>
<td>0.846</td>
<td>0.867</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>0.978</td>
<td></td>
</tr>
<tr>
<td>BR</td>
<td>1.00</td>
<td>0.844</td>
<td>0.874</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>0.977</td>
<td></td>
</tr>
<tr>
<td>TC</td>
<td>1.00</td>
<td>0.841</td>
<td>0.882</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>0.969</td>
<td></td>
</tr>
<tr>
<td>TC2</td>
<td>1.00</td>
<td>0.837</td>
<td>0.803</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>0.957</td>
<td></td>
</tr>
</tbody>
</table>

R, Reviewer; ES, equally spaced weighted counts; MR, mid-range weighted counts; BR, bottom-range weighted counts; TC, sum of polyp counts in all size categories; TC2, sum of polyp counts in the 2-4 mm and >4 mm categories.

*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, mid-range weighted counts, bottom-range weighted counts.
### TABLE 2. Variance components for video scoring*

<table>
<thead>
<tr>
<th>Polyp burden measure†</th>
<th>Patient-to-patient variance (relative variability)</th>
<th>Reviewer-to-reviewer variance (relative variability)</th>
<th>Residual variance (relative variability)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ES</td>
<td>797 (0.82)</td>
<td>14 (0.01)</td>
<td>162 (0.17)</td>
</tr>
<tr>
<td>MR</td>
<td>1689 (0.81)</td>
<td>22 (0.01)</td>
<td>387 (0.18)</td>
</tr>
<tr>
<td>BR</td>
<td>975 (0.81)</td>
<td>18 (0.01)</td>
<td>215 (0.18)</td>
</tr>
<tr>
<td>TC</td>
<td>243 (0.78)</td>
<td>14 (0.05)</td>
<td>53 (0.17)</td>
</tr>
<tr>
<td>TC2</td>
<td>97 (0.72)</td>
<td>3 (0.02)</td>
<td>34 (0.25)</td>
</tr>
</tbody>
</table>

ES, Equally spaced weighted counts; MR, mid-range weighted counts; BR, bottom-range weighted counts; TC, sum of polyp counts in all size categories; TC2, sum of polyp counts in the 2-4 mm and >4 mm categories.

*Variance components from linear mixed model analysis done on the various polyp burden measures. The model included variance components for patient-to-patient variance (biological variability), reviewer-to-reviewer variance (systematic reviewer variability), and residual variance (nonsystematic 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 patient-to-patient variability relative to the technical reviewer-related sources.

†See Methods section for the formulas for calculating the following polyp burden measures: equally spaced weighted counts, mid-range weighted counts, and bottom-range weighted counts.

### TABLE 3. Video scoring duration by reviewer

<table>
<thead>
<tr>
<th></th>
<th>Minimum</th>
<th>Q1*</th>
<th>Median</th>
<th>Q3*</th>
<th>Maximum</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Video length, sec†</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>73.0</td>
<td>137.3</td>
<td>203.5</td>
<td>277.3</td>
<td>384.0</td>
<td>212.5</td>
<td>89.2</td>
</tr>
<tr>
<td><strong>Normalized review time‡</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R1</td>
<td>1.1</td>
<td>1.6</td>
<td>2.1</td>
<td>3.7</td>
<td>7.4</td>
<td>2.8</td>
<td>1.5</td>
</tr>
<tr>
<td>R2</td>
<td>1.1</td>
<td>2.6</td>
<td>3.0</td>
<td>4.4</td>
<td>19.1</td>
<td>4.3</td>
<td>3.6</td>
</tr>
<tr>
<td>R3</td>
<td>2.1</td>
<td>3.1</td>
<td>3.8</td>
<td>4.4</td>
<td>14.1</td>
<td>4.5</td>
<td>2.6</td>
</tr>
<tr>
<td>All reviewers</td>
<td>1.1</td>
<td>2.3</td>
<td>3.2</td>
<td>4.2</td>
<td>19.1</td>
<td>3.9</td>
<td>2.8</td>
</tr>
</tbody>
</table>

Q1, First quartile; Q3, third quartile; SD, standard deviation; R, reviewer.

*For each measure, we summarize by the minimum, maximum, mean, median, first quartile, third quartile, and standard deviation.

†The video length is computed for each patient, 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 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.

### TABLE 4. Percent change in polyp burden by reviewer and measure

<table>
<thead>
<tr>
<th>Polyp burden measure*</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ES</td>
<td>−15.1% (−47.9% to 17.8%)†</td>
<td>−36.1% (−54.1% to −18.0%)</td>
<td>−36.2% (−48.1% to −24.3%)</td>
</tr>
<tr>
<td>MR</td>
<td>−15.1% (−51.6% to 21.5%)</td>
<td>−38.7% (−56.6% to −20.7%)</td>
<td>−38.9% (−50.9% to −27.0%)</td>
</tr>
<tr>
<td>BR</td>
<td>−13.3% (−51.7% to 25.2%)</td>
<td>−37.0% (−55.0% to −19.0%)</td>
<td>−37.0% (−49.0% to −24.9%)</td>
</tr>
<tr>
<td>TC</td>
<td>−15.0% (−39.0% to 10.2%)</td>
<td>−28.6% (−47.8% to −9.4%)</td>
<td>−30.7% (−43.5% to −17.8%)</td>
</tr>
<tr>
<td>TC2</td>
<td>−21.1% (−49.5% to 7.4%)</td>
<td>−39.5% (−56.7% to −22.2%)</td>
<td>−39.2% (−57.8% to −20.6%)</td>
</tr>
</tbody>
</table>

ES, Equally spaced weighted counts; MR, mid-range weighted counts; BR, bottom-range weighted counts; R1, reviewer 1; R2, reviewer 2; R3, reviewer 3; TC, sum of polyp counts in all size categories; TC2, sum of polyp counts in the 2-4 mm and >4 mm categories.

*See Methods section for formulas for calculating the following polyp burden measures: equally spaced weighted counts, mid-range weighted counts, and bottom-range weighted counts.

†Values are means (95% confidence interval).
effects accounting for another 15% to 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 the baseline and posttreatment polyp burden for a given patient. Because 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 nonsystematic 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 use 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, or real-time counting performed by the endoscopist. The photographs provide quantitative measurements but for only a small part of the colon, and they may vary depending on 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, reviewers blinded as to whether videos were from before or after treatment simply rated one video as better or worse than its paired mate. At that time, we had no convenient way to capture and review digitized videos. The video-based methods used in the current study reflect the ability to routinely capture, edit, maintain, and post for convenient Web-based review of large volumes of information. In the case of colonoscopy video, the approach used is both global and quantitative, so it has the potential to more accurately capture a

Figure 3. Cumulative distribution of percent changes in the total polyp burden after 6 months of celecoxib treatment by reviewer and polyp burden measure. In the top 3 panels, the 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 polyp burden = $(1 \times C_s) + (2 \times C_m) + (3 \times C_l)$; mid-range polyp burden = $(1 \times C_s) + (3 \times C_m) + (5 \times C_l)$; and bottom-range polyp burden = $(1 \times C_s) + (2 \times C_m) + (4 \times C_l)$. Total count indicates that the total polyp burden was determined by summing the counts for polyps in all 3 size groups; Count 2 indicates that total polyp burden was determined by summing the counts for medium (2-4 mm) and large (>4 mm) polyps only. ES, equally spaced; MR, mid-range; BR, bottom-range; TC, total count; TC2, count 2. $C_s$, $C_m$, and $C_l$ are the counts of small (<2 mm), medium (2-4 mm), and large (>4 mm) polyps.
patient’s total polyp burden. In our original celecoxib trial, 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 offices 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 by using selected photographs. However, treatment schedule, procedure, and patient population were very similar between the current study and our prior study. 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 the 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, raising the possibility that the assessment of response can be affected by the specialty 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. 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% to 85%; surgeon, 53% to 63%. These findings suggest that the proportion of patients with FAP who benefit from celecoxib might be higher than previously reported.

Although 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 proctocolectomy, use of existing and emerging chemopreventive agents can be expected to increase and to 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 nonacademic settings to not only post videos of colonoscopies but to assess the findings by using a rationally developed quantitative method for better communicating endoscopic results in FAP patients.

REFERENCES