Supplementary Materials

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 byt_{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 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, higher ρ indicates greater concordance, with $\rho=1$ indicating all raters gave the exact same rating to all videos, and $\rho=0.5$ indicating the variability across raters being equal in magnitude to the variability across videos.

In the latent model, for a score with five grades, a total of four 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 $(\gamma_{c-1}, \gamma_c]$, the observation is classified in to category c. The prior distributions were specified as follows: σ_ε^2 has an inverse-gamma prior, i.e., $1/\sigma_\varepsilon^2 \sim \text{Gamma}(1, 1)$, , and the category cutoffs $\gamma_c$are given independent uniform priors.

The posterior distributions of (σ_α^2, σ_ε^2, γ_c) were obtained using the MCMC algorithm . The concordance measure ρ for each posterior sample was calculated from σ_α^2 divided by σ_α^2+ σ_ε^2, from which the posterior mean and 95% posterior credible intervals were computed. Concordant measures should have $\rho>0.50$ at a minimum, since $\rho\leq0.5$ 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 $p=\text{Prob}(\rho\leq0.50|\text{data})$ as a measure of statistical significance, with $p<0.05$ indicating the level of agreement is significantly greater than this.

Simulation study

We performed a simulation study in order 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 data sets based on the different values of $\sigma_\alpha^2$and σ_ε^2 as follows:

- ICC = 0.5 implies $\sigma_\varepsilon^2= \sigma_\alpha^2$: the rater variation is the same as video variation
- ICC = 0.67 implies $\sigma_\varepsilon^2= \sigma_\alpha^2/2$: the rater variation is 1/2 of 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 p_c, obtain four cutoffs γ_c from Dirichlet distribution Dir(5, p_c)

4. Generate latent trait $t_{ij} = \alpha_i + \varepsilon_{ij}$ as follows:

   i. generateα_i from normal distribution $N(0, \sigma_\alpha^2)$ where $i = 1,\ldots, I$

   ii. givenα_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} using the cutoffs γ^j_c from Step 3

6. Estimate the posterior distribution of ρ using the MCMC algorithm [1, 2]

7. Claim a significant agreement if $\mathrm{Prob}(\rho \leq 0.5|\mathrm{data}) < p_L$, where p_L is disagreement parameter and should be set low such as 0.05 or 0.1.

Table 1 summarizes our simulation result with five different scenarios of ICC based on 24 raters and 24 videos 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. The simulation results showed that, a sample size of 24 raters and 24 videos will have at least 83% power for a concordance of ICC=0.70 based on this Bayesian multiple-rater modeling. More scenarios with 12 or 18 raters are shown in Table 2.

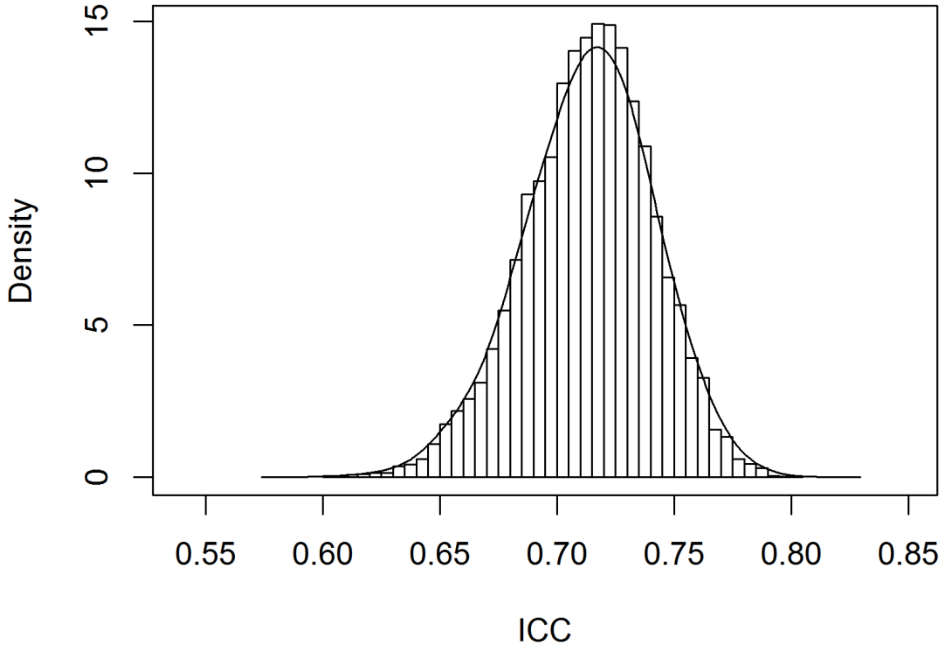
Supplemental Table 1: Power estimation based on number of raters (J=24) and number of videos (I = 24) from 100 simulated trials, assuming the distribution of ordered scores p_c are 0.30, 0.25, 0.20, 0.15, and 0.10, respectively. We claimed a significant agreement if Prob ($\rho \leq 0.5$ | data) <p_L:

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

Supplemental Table 2: Power estimation based on number of raters (J = 12 or 18) and number of 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 a significant agreement if Prob (ρ≤0.5 | data) <$p_{L:}$

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

Figure 1: Posterior distribution of ICC for IPSS score agreement based on 26 raters. The posterior mean (SE) of ICC is 0.710 (0.027) with 95% credible interval between 0.651 - 0.759



Supplemental Table 3: ICC for IPSS score for video ratings by demographic characteristics

Characteristics	Estimated value (s.e.)	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)
Gender		
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 or less (n=9)	0.743 (0.042)	(0.653, 0.819)
11 or more (n=17)	0.671 (0.038)	(0.594, 0.741)

*s.e=Standard Error; CI=Confidence Interval

Supplemental Table 4 : List of questions asked to the reviewers at the end of the reviews. (Please provide your opinion on the following statements

Sr. No	Question	Options	N (%)

1	The development of a staging system for colorectal polyposis will be helpful in communicating with colleagues regarding patient status	Strongly Agree
		Agree
		Neutral
		Disagree
		Strongly Disagree
			18(69)
			7(27)
			1(4)
			0
			0

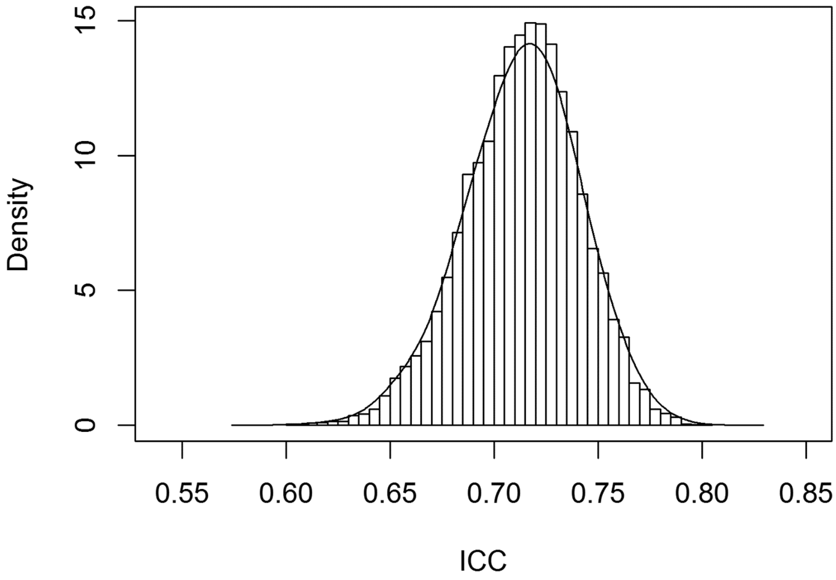
2	The development of a staging system for colorectal polyposis will be helpful in evaluating endpoints in clinical chemoprevention trials.	Strongly Agree
		Agree
		Neutral
		Disagree
		Strongly Disagree
			18(69)
			4(15.5)
			4(15.5)
			0
			0

3	Subject to my specific comments in the scoring sheet above, I am in general agreement with the present proposed IPSS	Strongly Agree
		Agree
		Neutral
		Disagree
		Strongly Disagree
			2(8)
			21(84)
			1(4)
			1(4)
			0

4	Subject to my comments in the scoring sheet above, I am in general agreement with the present proposed interventions by stage.	Strongly Agree
		Agree
		Neutral
		Disagree
		Strongly Disagree
			0
			16(62)
			8(30)
			2(8)
			0

Supplemental Table 5: Comment by reviewers on outlier cases.

Video No	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 Stage 1 and 2. Intervention hard to - 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 > 1cm. Perhaps offer an alternative stage category for <20 polyps, 1 or more >1cm, which may better fit this case." (Comment 2) "Tough case. Very few polyps but one 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 it 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 polyp, no >1cm, 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 (e.g. excision/revision) without histologic information. Suggest that it be pouch polyps 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 follo up would depend on that path result - as well as a few small raised areas elsewhere, although these are likely lymphoid aggregates." (Comment 2)


Density
0
5
10
15
0.55
0.60
0.65
0.70
0.75
0.80
0.85
ICC