Quantile regression enables reliable, uncertainty-aware endoscopic scoring in ulcerative colitis clinical trials

Sara Sangalli, Pooya Mobadersany, Shinobu Yamamoto, Chaitanya Parmar, Nicholas Skomrock, Tommaso Mansi, Oana Gabriela Cula, Kristopher Standish†, Pablo F. Damasceno†, Krishna Chaitanya†

Johnson & Johnson
kchaita6@its.jnj.com,salusanga@hotmail.com

Abstract. [1]
AI-based disease severity estimation in ulcerative colitis (UC) could substantially improve the detection of treatment effects in clinical trials. Their real-world deployment, however, is hampered by the lack of uncertainty models that can defer difficult cases to experts without eroding the treatment-effect sensitivity gains provided by AI scoring. Motivated by the weak labels and distributional shifts common in endoscopies from large, multi-site trials, we identify Quantile Regression (QR) as a principled strategy for selective regression. QR is distribution-free, yields interpretable uncertainty intervals, and accommodates non-Gaussian errors. We compare four single-model uncertainty approaches and ensembling on four large multicenter UC clinical-trial datasets. To our knowledge, this is the first study evaluating selective regression for trial-level decision making in UC trials. We evaluate selective behavior using MSE, RC curves, and trial-level treatment-effect size (Hedges' g) for dose selection and go/no-go decisions. Despite being a single model, QR consistently matches or competes with ensembles while reducing MSE relative to an uncertainty-agnostic baseline. Crucially, in a deployment-realistic dual-score deferral setting where uncertain AI scores are replaced by less treatment-sensitive expert scores, QR is the only method that increases Hedges' g over the baselines, showing that uncertainty-aware deferral preserves or improves AI treatment-effect sensitivity. These gains persist when up to 25% of cases are deferred, supporting QR as a reliable uncertainty-aware approach for disease-severity estimation in large-scale UC trials.

Keywords: Endoscopy · Uncertainty estimation · Reliable AI · Ulcerative Colitis (UC) · Quantile Regression (QR).

1 Introduction

In UC trials, clinical remission, typically defined using composite scores that include the Mayo Endoscopic Subscore (MES) [32], is the primary efficacy endpoint

[1] † Equal contribution as last authors.

on which dose selection and go/no-go decisions for advancing drug candidates ultimately rest. Endoscopy videos acquired in trials are graded using ordinal MES (0–3), assigned at the video level from the single worst severity colonic location, making it insensitive to subtle changes. Because grading is time consuming, costly, and subject to inter rater variability [2], trials rely on both local and central readers with adjudication, increasing cost and delaying readouts, particularly in large scale studies involving thousands of videos. This directly impacts the statistical power of the trial, motivating automated scoring that is accurate and sufficiently reliable for downstream decision making.

Recent efforts to automate MES grading [6] rely on ordinal labels from central readers. Because these labels are assigned at the video level, the task is weakly supervised and affected by class imbalance. To address this, newer approaches [7,31,25] shift to continuous MES (CMES) estimation, reframing the task as regression. CMES targets, defined as the mean of two expert scores, improve sensitivity to subtle changes, making automated estimation promising for reducing labeling burden and accelerating trials. However, endoscopic video analysis remains challenging, as videos are long and heterogeneous, many frames are ambiguous, and video-level labels carry reader noise. These factors give rise to complex error structures not captured by standard evaluation metrics. Moreover, because CMES is an averaged ordinal scoring, the associated uncertainty reflects the expected error of a continuous estimate rather than a calibrated probability over discrete grades, which is precisely the quantity required to decide whether a video should be scored automatically or deferred to an expert.

For clinical deployment, accuracy alone is insufficient. Models must identify potentially erroneous predictions and defer such cases to clinicians [5]. Selective regression [18,11], defined as abstaining when uncertainty exceeds a threshold, serves as a critical safety mechanism but remains under explored for continuous targets. Its evaluation requires metrics that jointly capture estimation error (MSE) and coverage, commonly visualized using Risk Coverage curves. Recent evidence shows that post-hoc selectors underperform methods with inherent uncertainty estimation, motivating uncertainty-aware models [17]. While prior work benchmarks selection on public datasets [13], its trade offs in large multi-site clinical trials remain unexplored. Crucially, the impact of selective scoring on trial level treatment effect estimation has not been addressed, despite its importance for dose selection and go/no-go decisions during drug development [12].

Motivated by the weak labels and distributional shifts inherent in multi-site UC trials, which lead standard regression models to produce unreliable and high variance predictions, we identify Quantile Regression (QR) [21,22] as a robust, distribution free strategy for selective CMES estimation. QR accommodates heteroscedastic and non-Gaussian errors and provides interpretable prediction intervals without parametric assumptions. These properties enable a selective abstention workflow that defers high uncertainty cases while preserving confident automated outputs. Rather than disentangling epistemic and aleatoric uncertainty, which is unreliable in multi-site clinical data, due to scale and complexity,

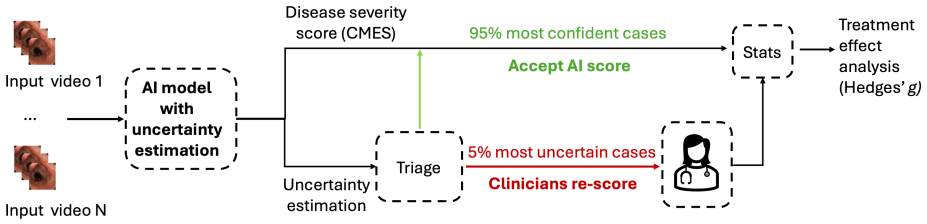

Fig. 1: Proposed uncertainty-aware deployment framework. For each video, an AI model predicts a CMES score with uncertainty; confident cases are accepted and uncertain ones are deferred to clinicians. Combined scores across deferral fractions (e.g., 5%) are used for treatment-effect estimation (Hedges' g).

we adopt a holistic deployment perspective where a single interval-based signal captures uncertainty and flags difficult cases to experts.

We evaluate QR against a comprehensive suite of uncertainty quantification methods, including MC Dropout [10], Deep Evidential Regression [3], Mean–Variance models [14], and Deep Ensembles [23], across four large multi-site UC trials. Our results show that QR is the only single model method that matches or exceeds ensemble performance at key operating points on Risk Coverage (RC) curve. Crucially, QR is also the only method whose deferral policy increases the trial-level treatment-effect size, measured by Hedges' g [15,16], beyond both manual MES and an uncertainty-agnostic CMES baseline [6], strengthening go/no-go and dose-selection decisions in settings where uncertain cases are deferred to clinicians. This links uncertainty estimation to trial-level decision making.

We identify and validate QR as an effective, distribution-free approach for selective UC disease severity estimation. Our contributions are: (1) to our knowledge, the first study to evaluate selective regression for trial-level decision making in UC trials, demonstrating that QR-based deferral of uncertain cases preserves and improves the treatment-effect sensitivity of AI scoring in a dual-score setting relevant for dose selection and go/no-go decisions; (2) a systematic comparison of uncertainty estimators under a deployment-realistic protocol, including RC analysis, validation to test threshold transfer, and multi-seed robustness across four large multi-site trials; and (3) a computationally efficient, clinically deployable workflow that reduces reader burden while supporting trial-level decisions.

2 Methods

2.1 Task definition

Given an endoscopic video $x \in \mathcal{X}$ from a multi-site UC clinical trial, represented as a variable-length sequence of T frames, $x = \{x_t\}_{t=1}^{T}$, the goal is to learn a regression model that estimates the continuous Mayo Endoscopic Subscore (CMES) $y \in [0, 3]$, defined as the average of MES scores from two expert raters. Continuous targets retain finer severity grades than the ordinal 0–3 scale and are

more sensitive to subtle mucosal change, but they also make prediction errors continuous valued, so a usable per video uncertainty estimate is essential for safe deployment. Beyond a point estimate $f_\theta(x) \in \mathbb{R}$, the model must output an uncertainty score $u_f(x) \in \mathbb{R}_{\geq 0}$ (e.g., estimated variance, interval width, or ensemble variance across models) that serves as a selection statistic for deferral to experts in uncertain cases.

2.2 Selective regression

Selective regression [18,11] lets a model abstain when uncertainty exceeds a threshold τ. Given a regression model $f_\theta : \mathcal{X} \to \mathbb{R}$ and uncertainty score $u_f : \mathcal{X} \to \mathbb{R}_{\geq 0}$, the selective predictor is:

$$\tilde{f}_\tau(x) = \begin{cases} f_\theta(x), & \text{if } u_f(x) \leq \tau, \\ \text{abstain}, & \text{otherwise.} \end{cases} \tag{1}$$

For a test set of N videos, coverage (Cov) and selective MSE at threshold τ are

$$\text{Cov}(\tau) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\{u_f(x_i) \leq \tau\}, \ \text{MSE}(\tau) = \frac{\sum_{i=1}^{N} \mathbb{I}\{u_f(x_i) \leq \tau\} \, (y_i - f_\theta(x_i))^2}{\sum_{i=1}^{N} \mathbb{I}\{u_f(x_i) \leq \tau\}}, \tag{2}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function and $MSE(\tau)$ is computed over the retained set. Plotting $\text{MSE}(\tau)$ against $\text{Cov}(\tau)$ yields the Risk–Coverage (RC) curve. When higher $u_f(x)$ correlates with larger error $|y - f_\theta(x)|$, abstention preferentially removes high error cases and lowers selective MSE, enabling deferral of difficult videos to experts. A useful uncertainty score must therefore rank videos by likely error so that reducing coverage removes high error cases efficiently. We also assess how thresholds selected on validation transfer to unseen test data (val→test). Next we describe the estimators used to produce u_f for the policy in Eq. (1).

2.3 Quantile regression (QR)

QR [21,22] estimates conditional quantiles of the CMES distribution, producing interpretable predictive intervals without parametric assumptions. We train a network to predict quantiles $\hat{q}_\alpha(x)$ at specified levels (e.g., $\alpha_L = 0.05, \ \alpha_U = 0.95$) using the Pinball loss [28]. For residual $r = y - \hat{q}_\alpha(x)$,

$$L_\alpha(r) = \begin{cases} \alpha \, r & \text{if } r \geq 0, \\ (\alpha - 1) \, r & \text{if } r < 0. \end{cases} \tag{3}$$

For two target quantiles, the total quantile loss is $L_{\alpha_{Tot}} = L_{\alpha_L}(y - \hat{q}_{\alpha_L}(x)) + L_{\alpha_U}(y - \hat{q}_{\alpha_U}(x))$. To encourage accurate point estimates, the model includes a third output $\mu(x)$ trained as the mean CMES regression via an L_2 penalty, giving $L_{\text{total}}(x, y) = \lambda \, L_{\alpha_{Tot}} + \text{MSELoss}(y, \mu(x))$, where λ is selected on the validation set. We define the **uncertainty score** as the interval width $u_f(x) =$

$\hat{q}_{\alpha_U}(x) - \hat{q}_{\alpha_L}(x)$, where wider intervals indicate greater uncertainty. As a distribution free method, QR captures input dependent uncertainty without a fixed noise model, accommodating heavy tailed errors and label noise with minimal computational overhead. This is well suited to multi-site trials. By estimating conditional quantiles rather than assuming homoscedastic Gaussian noise, QR adapts interval width to local prediction difficulty, widening intervals for ambiguous or out of distribution videos. Unlike sampling or ensemble based estimators, it requires a single forward pass and produces an uncertainty score that is directly interpretable for deferral and clinical triage.

2.4 Compared uncertainty methods

We compare five uncertainty estimators, each providing a score $u_f(x)$ that drives abstention (deferral) in Eq. 1.

(i) **Monte Carlo Dropout** [10] (MCD) performs M stochastic passes to obtain samples $\{f_\theta(x)^{(m)}\}_{m=1}^{M}$ and uses the sample variance as $u_f(x)$.

(ii) **Mean–Variance Gaussian models** [14] (MV) predict $\mu(x), \sigma^2(x)$ via Gaussian NLL and use $\sigma(x)$ as $u_f(x)$.

(iii) **Deep Evidential Regression** [3] (DER) predicts NIG parameters $(\mu, \nu, \alpha, \beta)$ and uses $u_f(x) = \sqrt{\mathrm{Var}_{\mathrm{pred}}(x)}$, where $\mathrm{Var}_{\mathrm{pred}} = \frac{\beta}{\alpha-1}\big(1 + \frac{1}{\nu}\big)$.

(iv) **Deep Ensembles** [23] (Ens.) aggregate M independently seeded models and use the standard deviation of their predictions as $u_f(x)$.

(v) We apply the same ensembling strategy to QR models (QR Ens.), combining the mean interval width and across-model standard deviation as $u_f(x)$ to assess whether ensembling improves upon a single QR model.

MCD and Ens. primarily capture epistemic uncertainty, while MV, DER, and QR capture aleatoric uncertainty, although this distinction often blurs in large-scale, real-world settings, as shown in prior evaluations [19,20,17].

3 Experiments

Dataset. For model development, we use four multi-site UC clinical trial datasets (VEGA[9], JAKUC[29], UNIFI[30] and ASTRO[24]), totaling 5,918 endoscopic videos. The score distribution is naturally skewed: screening-stage enrollment over-represents moderate to severe disease (MES 2–3) and includes fewer normal to mild cases (MES 0–1; e.g., 170 of 809 videos in unseen ANTHEM-UC[1] trial), which we relate to estimated uncertainty in Sec. 4. Data were split at the patient and video level into 60%/20%/20% train/validation/test. Evaluation uses four test sets: 20% hold-out splits from Trials UNIFI and ASTRO (in distribution) and two fully unseen trials QUASARPh2b[27] and ANTHEM-UC[1] (to assess generalization under distribution shift). The UNIFI and ASTRO hold-out test sets contain 153 (∼650k frames) and 587 videos (∼7.1M frames), while unseen trials QUASARPh2b and ANTHEM-UC contain 614 (∼14M frames) and 809 videos (>21M frames).

Baseline regression model. Following ARGES [6], we use a ViT-B backbone [8] pretrained via DINOv2 [26] on ∼61M endoscopic frames for frame-level feature extraction. These features are processed by a transformer block for temporal modeling, followed by an attention-based multiple-instance learning layer that aggregates frame-level representations into a video-level embedding. A method-specific prediction head is then used to produce the final video-level output. The CMES estimate is therefore produced directly from the aggregated video-level embedding rather than by averaging frame-level predictions or uncertainty-weighted frame scores. Models are trained for 30 epochs using MSE loss, weighted sampling, a learning rate of 10^{-4}, and weight decay of 10^{-5}. Dropout rates of 0.25 and 0.5 are applied in the transformer layers and final prediction head, respectively.

Uncertainty extraction. For each method (Sec. 2), we compute a per-video uncertainty score $u_f(x)$, using validation set for threshold selection and test set for evaluation. For QR, model predicts two quantiles ($\alpha_L = 0.05$, $\alpha_U = 0.95$) and a CMES estimate μ, trained with Pinball and MSE losses; interval width defines $u_f(x)$. MCD uses $T = 40$ stochastic passes with dropout rates of 0.25 and 0.5 retained in the transformer layers and prediction head, respectively, and computes sample variance. MV models predict $\mu(x)$ and $\log \sigma^2(x)$ via Gaussian NLL, using $\sigma(x)$ as uncertainty. DER predicts NIG parameters, with $u_f(x)$ as square root of predictive variance. Ensembles use $M = 5$ independently initialized models and define uncertainty as prediction standard deviation. We also evaluate a QR ensemble (QR Ens), where uncertainty is the sum of mean QR interval width and ensemble standard deviation, to assess whether ensembling improves over a single QR model.

Evaluation protocol. We rank test samples by $u_f(x)$ to generate RC curves with MSE as risk, and report selective MSE at fixed operating points. **Calibration coverage.** For QR, we additionally evaluate interval calibration. Given lower and upper quantile predictions $(q_{\alpha_L}, q_{\alpha_U})$, calibration coverage is computed as the proportion of samples whose ground-truth CMES score falls within the predicted interval. For a well-calibrated model, the empirical coverage should closely match the nominal coverage level. Ground-truth labels are the mean of two readers, except for ANTHEM-UC trial (single rater available). We report consistent coverage levels (0.8, 0.9) across experiments for direct comparison between RC curves and tabulated operating points. For threshold generalization (val→test), we select τ on validation and apply it to test set, reporting differences on ANTHEM-UC, largest unseen trial. Treatment effects on ANTHEM-UC are computed using per-subject change scores $\Delta = \text{score}_{\text{Week12}} - \text{score}_{\text{Week0}}$, where more negative values indicate greater clinical improvement. Hedges' g is computed under the evaluated scoring scheme: manual MES, AI CMES, or hybrid AI CMES+human deferral. For each active dose, we compare against placebo using Hedges' g [15]. With placebo/treatment means, variances, and sample sizes $(\bar{X}_{\text{P}}, s^2_{\text{P}}, n_{\text{P}})$ and $(\bar{X}_{\text{T}}, s^2_{\text{T}}, n_{\text{T}})$, the pooled SD is $S_p = \sqrt{\frac{(n_{\text{P}}-1)s^2_{\text{P}} + (n_{\text{T}}-1)s^2_{\text{T}}}{n_{\text{P}}+n_{\text{T}}-2}}$, and $g = J(\bar{X}_{\text{P}} - \bar{X}_{\text{T}})/S_p$, where J is the small-sample correction. Hedges' g is a bias-corrected standardized mean difference commonly used to quantify treatment–

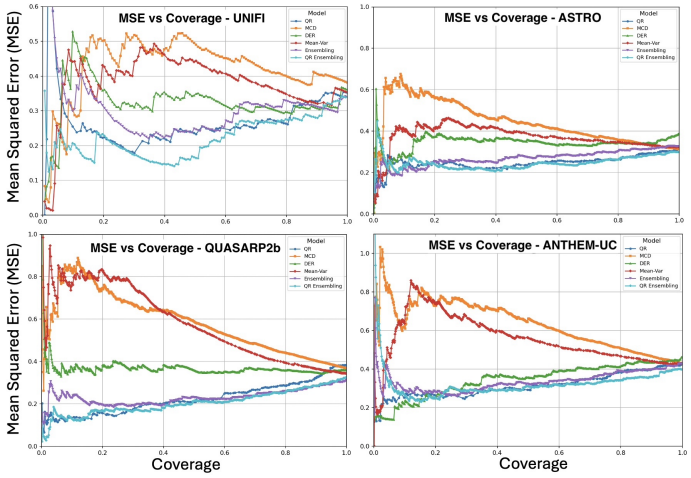

Fig. 2: Risk–coverage curves showing selective MSE across coverage levels. QR and QR Ensembles achieve the lowest MSE at most coverage levels.

placebo separation in trials, supporting dose selection and go/no-go decisions [7]; positive g reflects greater improvement over placebo. Because g depends on both mean separation $(\bar{X}_{\mathrm{P}} - \bar{X}_{\mathrm{T}})$ and pooled std dev S_p, we examine whether observed gains are accompanied by preserved or improved treatment–placebo separation rather than variance reduction alone. We further test robustness under deployment by replacing the most uncertain AI scores with ordinal expert MES scores (Fig. 1, Table 2). As a control experiment, we also replace the same proportion of AI predictions selected uniformly at random, allowing us to determine whether improvements stem from uncertainty-guided identification of challenging cases rather than score replacement alone. All experiments are repeated with multiple seeds and folds; we report mean and standard deviation.

4 Results

Fig. 2 shows RC curves for each trial. QR is comparable to QR Ens. and consistently attains the lowest or second-lowest selective MSE across most coverage levels. Table 1 reports representative operating points: full coverage (standard MSE, with ARGES baseline) and reduced-coverage settings quantifying selective gains. QR also exhibits slightly lower MSE than the ARGES baseline at full coverage. While both methods share the same backbone and video aggregation architecture, the additional quantile objectives may provide a regularizing effect that encourages more robust feature learning and modestly improves CMES estimation accuracy. UNIFI test set elevated low-coverage variability reflects its small 20% hold-out set (153 videos), where retaining only the most confident samples leaves few cases; this effect disappears on the larger unseen trials QUASARPh2b and ANTHEM-UC. Across seeds and folds, QR is the most

stable estimator (standard deviation ≈ 0.01 vs. 0.02 for the next-best method). We further verified via calibration-coverage [4] that QR attains 90–95% coverage across datasets, indicating well-calibrated intervals aligned with empirical uncertainty.

The val→test experiment (Fig. 3) evaluates whether validation-selected thresholds generalize to the largest unseen ANTHEM-UC trial. For each method, we select the threshold at target validation coverage (0.9, 0.8, 0.7), apply it to the test set, and compute $\Delta\text{Cov}_{\text{val}-\text{test}}$ and $\Delta\text{MSE}_{\text{val}-\text{test}}$; points nearer the origin indicate better transfer. QR shows smaller shifts than most methods and performs comparably to ensembles, indicating reliable generalization.

Among the top 10% most uncertain ANTHEM-UC trial videos deferred by QR (81 of 809), 43 belong to the under-represented normal/mild disease classes (MES 0–1), reflecting class imbalance, while the remaining cases correspond to moderate-to-severe disease (MES 2–3) and are often associated with low video quality or ambiguity near ordinal score boundaries. This suggests QR interval width tracks predictive difficulty rather than label frequency alone. Since these factors mix aleatoric and epistemic sources that are difficult to separate at real-world scale, we use interval width as a holistic deferral criterion. While practical, this formulation does not explicitly disentangle aleatoric and epistemic uncertainty. Additionally, our expert-deferral analysis is retrospective and simulated rather than prospectively evaluated in a clinical workflow, which will be investigated in future studies.

In Table 2, we quantify drug–placebo separation on unseen ANTHEM-UC trial using Hedges' g, comparing QR against human MES, ARGES CMES baseline, and deep ensembles, the most competitive alternative. QR attains the highest g for both doses and is the only single-model method to improve over the uncertainty-agnostic ARGES baseline; asterisks denote significance by Student's t-test [33]. In a simulated deployment, replacing the top 5%, 15%, and 25% most uncertain CMES predictions with human MES preserves and generally improves treatment-effect separation, peaking at 15% deferral. In contrast, randomly replacing the same fraction of predictions consistently yields lower Hedges' g values for both dose groups. These results indicate that the gains arise from identifying genuinely challenging cases for expert review rather than simply mixing AI and human scores. Importantly, the uncertainty-based deferral strategy maintains treatment–placebo separation well above the manual-MES baseline even at higher deferral rates and supports clearer differentiation between doses, potentially facilitating dose selection (200mg over 400mg). Furthermore, the gains are unlikely to arise from a simple smoothing or variance-reduction effect: if improvement were driven primarily by variance reduction, replacing uncertain AI scores with discrete expert MES, particularly at random, would be expected to move the effect size toward the MES baseline. Instead, uncertainty-based deferral preserves or improves Hedges' g and consistently outperforms random deferral, suggesting that QR uncertainty estimates capture clinically meaningful prediction uncertainty and can be used to selectively route difficult cases for expert assessment.

Table 1: Operating points from Fig. 2. MSE at full coverage (1.0) and reduced coverage levels (0.9, 0.8) on the test set.

Trials:	UNIFI	ASTRO	QUASAR Ph2b	ANTHEM -UC
Method	**MSE (↓) on coverage 1.0**			
ARGES	$.409_{\pm.05}$	$.409_{\pm.04}$	$.335_{\pm.04}$	$.486_{\pm.05}$
MCD	$.382_{\pm.04}$	$\underline{.313}_{\pm.04}$	$.369_{\pm.05}$	$\underline{.427}_{\pm.05}$
MV	$.354_{\pm.05}$	$.324_{\pm.04}$	$.343_{\pm.04}$	$.430_{\pm.05}$
DER	$.361_{\pm.06}$	$.387_{\pm.05}$	$.358_{\pm.05}$	$.461_{\pm.06}$
Ens.	$\mathbf{.340}_{\pm.02}$	$.327_{\pm.02}$	$\mathbf{.315}_{\pm.02}$	$\underline{.427}_{\pm.03}$
QR Ens.	$\mathbf{.340}_{\pm.01}$	$\mathbf{.299}_{\pm.01}$	$\underline{.326}_{\pm.01}$	$\mathbf{.399}_{\pm.02}$
QR	$\underline{.351}_{\pm.01}$	$.316_{\pm.01}$	$.383_{\pm.01}$	$.430_{\pm.01}$
	MSE (↓) at coverage 0.9			
MCD	$.407_{\pm.05}$	$.335_{\pm.04}$	$.404_{\pm.06}$	$.466_{\pm.06}$
MV	$.321_{\pm.05}$	$.333_{\pm.04}$	$.365_{\pm.05}$	$.440_{\pm.05}$
DER	$\underline{.308}_{\pm.06}$	$.345_{\pm.05}$	$.348_{\pm.05}$	$.434_{\pm.06}$
Ens.	$.312_{\pm.03}$	$.313_{\pm.02}$	$\mathbf{.284}_{\pm.02}$	$.399_{\pm.03}$
QR Ens.	$\mathbf{.280}_{\pm.01}$	$\mathbf{.285}_{\pm.01}$	$\underline{.285}_{\pm.02}$	$\mathbf{.359}_{\pm.02}$
QR	$.327_{\pm.01}$	$\underline{.293}_{\pm.01}$	$.323_{\pm.01}$	$\underline{.399}_{\pm.01}$
	MSE (↓) at coverage 0.8			
MCD	$.400_{\pm.05}$	$.358_{\pm.05}$	$.442_{\pm.06}$	$.507_{\pm.06}$
MV	$.348_{\pm.05}$	$.348_{\pm.04}$	$.394_{\pm.05}$	$.466_{\pm.05}$
DER	$.297_{\pm.06}$	$.346_{\pm.05}$	$.359_{\pm.05}$	$.416_{\pm.06}$
Ens.	$.329_{\pm.03}$	$.303_{\pm.02}$	$\underline{.261}_{\pm.02}$	$.390_{\pm.03}$
QR Ens.	$\underline{.275}_{\pm.01}$	$\mathbf{.262}_{\pm.01}$	$\mathbf{.253}_{\pm.02}$	$\mathbf{.343}_{\pm.01}$
QR	$\mathbf{.263}_{\pm.01}$	$\underline{.264}_{\pm.01}$	$.287_{\pm.02}$	$\underline{.382}_{\pm.02}$

Fig. 3: Validation-to-test differences in coverage and MSE using uncertainty thresholds selected on the validation set at target coverage levels. Results are shown for unseen ANTHEM-UC trial.

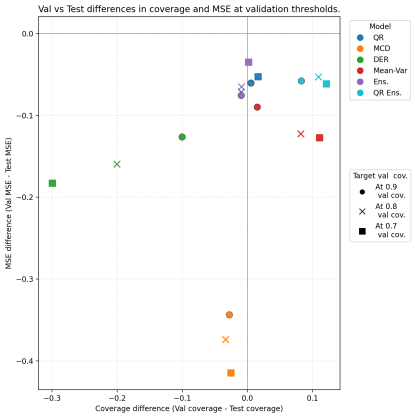

Table 2: Hedges' g (↑) on the unseen ANTHEM-UC trial, comparing treatment doseages (200mg and 400mg) versus placebo (Pbo). QR improves treatment–placebo separation over human MES, ARGES, and ensemble baselines. In a simulated deployment, the top $x\%$ most uncertain QR predictions are replaced with human MES (uncertainty-based deferral). For comparison, the same proportion of predictions is replaced at random (random deferral). Uncertainty-based deferral preserves or improves treatment-effect separation, whereas random deferral does not maintain these gains beyond 5% deferral ($*$ indicates statistical significance).

	Hedges' g (↑)					**Uncertainty-based Deferral (QR)**			**Random Deferral**		
Dose	**MES**	**ARGES**	**Ens.**	**QR**	**Ens.**	**5%**	**15%**	**25%**	**5%**	**15%**	**25%**
200mg VS Pbo	.55	.734*	.706*	.71*	**.742***	.768*	.81*	.733*	.723*	.661	.607
400mg VS Pbo	.53	.659*	.643*	.656*	**.667***	.650*	.769*	.710*	.632*	.625	.640

5 Conclusion

We advocate QR as a practical, distribution-free uncertainty method for selective CMES estimation. Across four clinical trials, QR matches multi-model ensembles on RC curves, reduces selective MSE versus an uncertainty-agnostic

CMES baseline, and is the only method that increases Hedges' g, supporting go/no-go and dose selection decisions. Its val→test threshold transfer is reliable, and gains persist under a human-in-the-loop deferral policy. We treat uncertainty holistically rather than explicitly disentangling uncertainty sources, linking high uncertainty to data imbalance and video difficulty in real-world trial data. Overall, QR provides uncertainty estimates that support robust trial-level decision making in deployment settings.

Disclosure of Interests: Sara Sangalli contributed to this work during an internship at Johnson & Johnson. All other authors were employees of Johnson & Johnson at the time this research was conducted and may own Johnson & Johnson stock or stock options.

References

1. Abreu, M., Siegmund, B., Surace, L., et al.: Icotrokinra, a targeted oral peptide that selectively blocks il-23 receptor activation, in moderately to severely active ulcerative colitis: week 12 results from the phase 2b, randomized, double-blind, placebo-controlled, treat-through, dose-ranging anthemuc trial. Oral presentation OP206 at United European Gastroenterology Week (UEGW) (2025)
2. Agbelese, D., Chaitanya, K., Pati, P., Parmar, C., Mobadersany, P., Fadnavis, S., Surace, L., Yarandi, S., Ghanem, L.R., Lucas, M., Mansi, T., Cula, G.O., Damasceno, P.F., Standish, K.: Megan: Mixture of experts for robust uncertainty estimation in endoscopy videos. In: Sudre, C.H., Hoque, M.I., Mehta, R., Ouyang, C., Qin, C., Rakic, M., Wells, W.M. (eds.) Uncertainty for Safe Utilization of Machine Learning in Medical Imaging. pp. 3–13. Springer Nature Switzerland, Cham (2026)
3. Amini, A., Schwarting, W., Soleimany, A., Rus, D.: Deep evidential regression. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20, Curran Associates Inc., Red Hook, NY, USA (2020)
4. Angelopoulos, A.N., Bates, S.: A gentle introduction to conformal prediction and distribution-free uncertainty quantification (2022), https://arxiv.org/abs/2107.07511
5. Azad, T.D., Krumholz, H.M., Saria, S.: Principles to guide clinical ai readiness and move from benchmarks to real-world evaluation. Nature Medicine (2026). https://doi.org/10.1038/s41591-025-04198-1, published online January 23, 2026
6. Chaitanya, K., Damasceno, P.F., Fadnavis, S., Mobadersany, P., Parmar, C., Scherer, E., Zemlianskaia, N., Surace, L., Ghanem, L.R., Cula, O.G., et al.: Arges: Spatio-temporal transformer for ulcerative colitis severity assessment in endoscopy videos. In: International Workshop on Machine Learning in Medical Imaging. pp. 201–211. Springer (2024)
7. Chaitanya, K., Scherer, E., Damasceno, P.F., Cerna, A.U., Zemlianskaia, N., Mobadersany, P., Parmar, C., Skomrock, N., Surace, L., Yarandi, S., et al.: Tu1992 arges-cmes: A novel, continous scoring for ulcerative colitis disease severity: Association with clinical measures and treatment response variables. Gastroenterology **166**(5), S–1482 (2024)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2021), https://arxiv.org/abs/2010.11929
9. Feagan, B.G., Sands, B.E., Sandborn, W.J., Germinaro, M., Vetter, M., Shao, J., Sheng, S., Johanns, J., Panés, J., Tkachev, A., et al.: Guselkumab plus golimumab combination therapy versus guselkumab or golimumab monotherapy in patients with ulcerative colitis (vega): a randomised, double-blind, controlled, phase 2, proof-of-concept trial. The Lancet Gastroenterology & Hepatology **8**(4), 307–320 (2023)
10. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: Balcan, M.F., Weinberger, K.Q. (eds.) Proceedings of The 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 48, pp. 1050–1059. PMLR, New York, New York, USA (20–22 Jun 2016), https://proceedings.mlr.press/v48/gal16.html
11. Geifman, Y., El-Yaniv, R.: Selective classification for deep neural networks. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 4885–4894. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)

12. Götte, H., Kirchner, M., Sailer, M.O., Kieser, M.: Optimal designs for phase II/III drug development programs including methods for discounting of phase II results. BMC Medical Research Methodology **20**(1), 253 (2020). https://doi.org/10.1186/s12874-020-01093-w
13. Gustafsson, F., Danelljan, M., Schön, T.: How reliable is your regression model's uncertainty under real-world distribution shifts? Transactions on Machine Learning Research (02 2023)
14. Gustafsson, F.K., Danelljan, M., Schon, T.B.: Evaluating scalable bayesian deep learning methods for robust computer vision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2020)
15. Hedges, L.V.: Estimation of effect size from a series of independent experiments. Psychological Bulletin **92**(2), 490–499 (1982). https://doi.org/10.1037/0033-2909.92.2.490
16. Hedges, L.V.: Effect sizes in cluster-randomized designs. Journal of Educational and Behavioral Statistics **32**(4), 341–370 (2007). https://doi.org/10.3102/1076998606298043, https://doi.org/10.3102/1076998606298043
17. Jaeger, P.F., Lüth, C.T., Klein, L., Bungert, T.J.: A call to reflect on evaluation practices for failure detection in image classification. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=YnkGMIh0gvX
18. Jiang, W., Zhao, Y., Wang, Z.: Risk-controlled selective prediction for regression deep neural network models. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (2020). https://doi.org/10.1109/IJCNN48605.2020.9207676
19. Jürgens, M., Meinert, N., Bengs, V., Hüllermeier, E., Waegeman, W.: Is epistemic uncertainty faithfully represented by evidential deep learning methods? In: Proceedings of the 41st International Conference on Machine Learning. ICML'24, JMLR.org (2024)
20. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 5580–5590. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
21. Koenker, R., Bassett, G.: Regression quantiles. Econometrica **46**(1), 33–50 (1978), http://www.jstor.org/stable/1913643
22. Koenker, R., Hallock, K.F.: Quantile regression. The Journal of Economic Perspectives **15**(4), 143–156 (2001), http://www.jstor.org/stable/2696522
23. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 6405–6416. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
24. Long, M., Allegretti, J.R., Danese, S., Germinaro, M., Baker, T., Alvarez, Y., Kavalam, M., Han, C., Jörgens, S., Jiang, L., et al.: Efficacy and safety of subcutaneous guselkumab induction therapy in participants with moderately to severely active ulcerative colitis (astro): a double-blind, treat-through, randomised, placebo-controlled, phase 3 trial. The Lancet Gastroenterology & Hepatology **11**(4), 284–298 (2026)
25. Mobadersany, P., Chaitanya, K., Parmar, C., Halchenko, Y., Yamamoto, S., Surace, L., Yarandi, S., Skomrock, N., Ghanem, L.R., Cula, G.O., et al.: 169 arges-uc: Ai-based continuous scoring for disease severity in ulcerative colitis enhances detection

of treatment effects and reduces sample size requirements in clinical trials for measuring endoscopic treatment differences. Gastrointestinal Endoscopy **103**(5), S–869 (2026)
26. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2024), https://arxiv.org/abs/2304.07193
27. Peyrin-Biroulet, L., Allegretti, J.R., Rubin, D.T., Bressler, B., Germinaro, M., Huang, K.H.G., Shipitofsky, N., Zhang, H., Wilson, R., Han, C., et al.: Guselkumab in patients with moderately to severely active ulcerative colitis: Quasar phase 2b induction study. Gastroenterology **165**(6), 1443–1457 (2023)
28. Romano, Y., Patterson, E., Candes, E.: Conformalized quantile regression. In: Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019)
29. Sands, B.E., Sandborn, W.J., Feagan, B.G., Lichtenstein, G.R., Zhang, H., Strauss, R., Szapary, P., Johanns, J., Panes, J., Vermeire, S., et al.: Peficitinib, an oral janus kinase inhibitor, in moderate-to-severe ulcerative colitis: results from a randomised, phase 2 study. Journal of Crohn's and Colitis **12**(10), 1158–1169 (2018)
30. Sands, B.E., Sandborn, W.J., Panaccione, R., O'Brien, C.D., Zhang, H., Johanns, J., Adedokun, O.J., Li, K., Peyrin-Biroulet, L., Van Assche, G., et al.: Ustekinumab as induction and maintenance therapy for ulcerative colitis. New England Journal of Medicine **381**(13), 1201–1214 (2019)
31. Scherer, E., Chaitanya, K., Damasceno, P.F., Cerna, A.U., Zemlianskaia, N., Mobadersany, P., Parmar, C., Skomrock, N., Surace, L., Yarandi, S., et al.: 544 arges-cmes: A novel, continuous scoring for ulcerative colitis disease severity: Prediction and sensitivity in assessing treatment response. Gastroenterology **166**(5), S–125 (2024)
32. Schroeder, K.W., Tremaine, W.J., Ilstrup, D.M.: Coated oral 5-aminosalicylic acid therapy for mildly to moderately active ulcerative colitis. New England Journal of Medicine **317**(26), 1625–1629 (1987)
33. Student: The probable error of a mean. Biometrika **6**(1), 1–25 (1908), http://www.jstor.org/stable/2331554