AP&T Alimentary Pharmacology & Therapeutics   WILEY

Agreement between local and central reading of endoscopic disease activity in ulcerative colitis: results from the tofacitinib OCTAVE trials

Brian G. Feagan[1] 🔵 | Reena Khanna[2] | William J. Sandborn[3] | Séverine Vermeire[4] |
Walter Reinisch[5] 🔵 | Chinyu Su[6] | Leonardo Salese[6] | Haiyun Fan[6] |
Jerome Paulissen[7] | Deborah A. Woodworth[6] | Wojciech Niezychowski[6] |
Bruce E. Sands[8]

[1]Robarts Research Institute, Western University, London, Ontario, Canada

[2]Department of Medicine, Division of Gastroenterology, Western University, London, Ontario, Canada

[3]Division of Gastroenterology, University of California San Diego, La Jolla, CA, USA

[4]Department of Gastroenterology, University Hospitals Leuven, Leuven, Belgium

[5]Department of Internal Medicine III, Medical University of Vienna, Vienna, Austria

[6]Pfizer Inc, Collegeville, PA, USA

[7]Pfizer Inc, New York, NY, USA

[8]Dr. Henry D. Janowitz Division of Gastroenterology, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Correspondence
Leonardo Salese, Pfizer Inc, 500 Arcola Road, Collegeville, PA 19426, USA.
Email: Leonardo.Salese@pfizer.com

Summary

Background: Endoscopy is routine in trials of ulcerative colitis therapies.

Aim: To investigate agreement between central and local Mayo endoscopic subscore (MES) reads in the OCTAVE programme

Methods: Flexible sigmoidoscopy was performed in tofacitinib induction (OCTAVE Induction 1&2, NCT01465763 and NCT01458951), maintenance (OCTAVE Sustain, NCT01458574) and open-label, long-term extension (OCTAVE Open, NCT01470612) studies. Kappa statistics and Bowker's tests evaluated agreement/disagreement between centrally and locally read MES, with potential determinants of differences analysed by logistic regression.

Results: Moderate-to-substantial agreement was observed between central and local reads at screening (77.1% agreement; kappa 0.62 [95% confidence interval 0.59-0.66]), OCTAVE Induction 1&2 week (Wk) 8 (63.8%; 0.62 [0.59-0.66]), OCTAVE Sustain Wk 52 (55.6%; 0.56 [0.50-0.62]) and for induction non-responders at OCTAVE Open month 2 (59.9%; 0.54 [0.48-0.60]). Where disagreements occurred, local reads were systematically lower than central reads at OCTAVE Induction 1&2 Wk 8, OCTAVE Sustain Wk 52 and OCTAVE Open month 2 (Bowker's $P < 0.0001$); this difference was not observed at screening ($P = 0.0852$). Using multivariable logistic regression, geographical region, C-reactive protein (Wk 8), partial Mayo score (Wk 8) and prior tumour necrosis factor antagonist failure were associated with disparity at OCTAVE Induction 1&2 Wk 8 ($P < 0.05$). In OCTAVE Induction 1&2 and OCTAVE Sustain, significantly higher proportions of patients endoscopic improvement, remission and endoscopic remission with tofacitinib vs placebo, using either central or local reads.

The Handling Editor for this article was Dr Nicholas Kennedy, and it was accepted for publication after full peer-review.

- -

Conclusion: Moderate-to-substantial agreement was observed between central and local endoscopic reads. Where disagreements occurred, local reads were systematically lower than central reads at most timepoints, suggesting potential bias. ClinicalTrials.gov identifier: NCT01465763, NCT01458951, NCT01458574, NCT01470612.

1 | INTRODUCTION

Endoscopy is the measure of disease activity most commonly used in ulcerative colitis (UC)[1] clinical trials for determination of patient eligibility and evaluation of efficacy.[2] Although considerable progress has been made towards validating endoscopic scoring, continued efforts are needed to optimise the sensitivity and reproducibility of endoscopic indices for the detection of treatment effects. Although endoscopic scoring by site (local) readers is convenient, a mesalamine induction study demonstrated that this may lead to biased results, higher placebo rates and diminished sensitivity for detection of a treatment effect relative to central reading.[3] Although causes of systematic disagreement between local and central readers are poorly understood, the former may be influenced by knowledge of the patient's clinical presentation, aspired outcomes for therapy and the chronology of the treatment course and sequence of visits within a study protocol. To limit risk of bias in endoscopic assessment, regulatory agencies (the US Food and Drug Administration and the European Medicines Agency) recommend that assessment of endoscopic disease activity be performed by central reading.[4,5] However, optimal methodology for central reading has not been thoroughly investigated.

The Mayo endoscopic subscore (MES) is a component of the Mayo score, and is recommended by major regulatory bodies for both eligibility and efficacy assessments in UC clinical trials.[4,5] It consists of a 4-point scoring system; scores range from 0 to 3, with higher MES indicating more severe disease activity (0 = normal or inactive disease; 1 = mild disease [erythema, decreased vascular pattern]; 2 = moderate disease [marked erythema, absent vascular pattern, any friability, erosions]; 3 = severe disease [spontaneous bleeding, ulceration]). Generally, induction trials require a minimum MES of 2 for eligibility.[6–8] Endoscopic improvement, the endoscopic component of the clinical remission definition that is conventional for registration studies, is defined as a MES of 0 or 1[4]; endoscopic remission is defined as a MES of 0. A 1-point improvement in the MES is accepted as a clinically meaningful change and is referred to as endoscopic response. Large variability or bias in reading the MES can negatively affect remission and response estimates,[3] and is, therefore, an important consideration in UC trial design.

Tofacitinib is an oral, small molecule Janus kinase inhibitor for the treatment of UC. The efficacy and safety of tofacitinib was established for the treatment of moderate-to-severe UC in a phase 2 induction study[9] and the phase 3 OCTAVE programme, which comprised: two 8-week, randomised, double-blind, placebo-controlled induction studies (OCTAVE Induction 1 and 2, NCT01465763 and NCT01458951); a

52-week, randomised, double-blind, placebo-controlled maintenance study (OCTAVE Sustain, NCT01458574)[7]; and an open-label, long-term extension study (OCTAVE Open, NCT01470612).[10] OCTAVE was among the first large, phase 3 programmes in UC that used central reading of the MES. Here, we investigated agreement and potential sources of disagreement between central and local endoscopic reads in the OCTAVE clinical programme. Efficacy endpoints, based on centrally and locally read MES, were also assessed.

2 | METHODS

2.1 | Patients and study design

Patients had moderately to severely active UC, defined by a total Mayo score of ≥6, with a Mayo rectal bleeding subscore of ≥1 and a MES of ≥2 (centrally read), and had failed or were intolerant to treatment with oral or intravenous corticosteroids, azathioprine or 6-mercaptopurine, or tumour necrosis factor (TNF) antagonists. Full details of permitted and prohibited concomitant medications, and corticosteroid tapering, in the OCTAVE clinical programme are provided in the Supporting Information. A study design overview is provided in Figure 1 and the Supporting Information; full details have been reported previously.[7]

Endoscopic scores from central reads were used for the primary efficacy analyses in OCTAVE Induction 1 and 2, and OCTAVE Sustain. Endoscopic improvement (referred to as mucosal healing in the OCTAVE protocols) was defined as a MES of ≤1. Remission was defined as a total Mayo score of ≤2 with no individual subscore >1, and a rectal bleeding subscore of 0. Endoscopic remission was defined as a MES of 0.

2.2 | Endoscopic assessment

In the OCTAVE clinical programme, flexible sigmoidoscopy with MES evaluation was performed at various time points, including at the induction screening visit, end of induction, midway through maintenance, end of maintenance and at Month 2 of OCTAVE Open. For induction screening, colonoscopy was performed instead of sigmoidoscopy for patients at risk of colorectal cancer. In OCTAVE Open, the MES for induction non-responders was assessed based on both central and local reads at Month 2.

Per protocol, MES was scored as described in the Supporting Information. Central endoscopic reading was used to determine eligibility for entry into OCTAVE Induction 1 and 2, and progression into OCTAVE Sustain, to qualify patients for early withdrawal
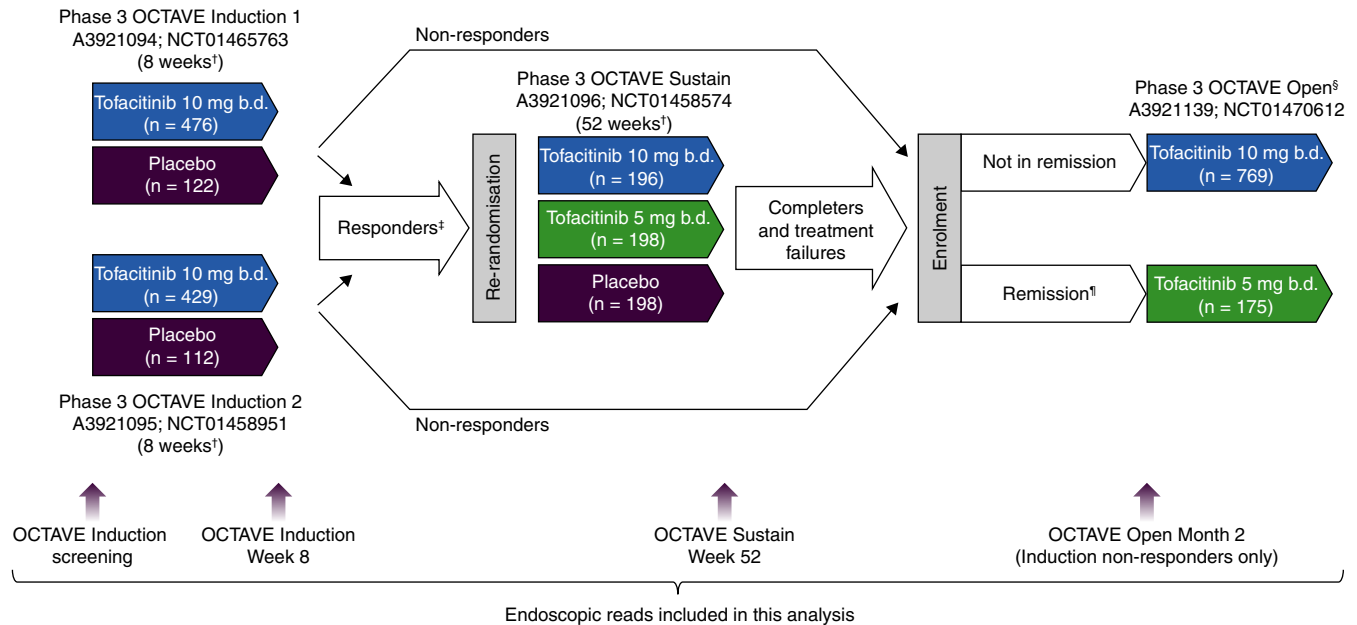
FIGURE 1 Overview of the tofacitinib phase 3 OCTAVE clinical programme. b.d., twice daily; n, number of patients treated. [†]Final complete efficacy assessment at Week 8/52. Treatment continued up to Week 9/53. [‡]Clinical response in OCTAVE Induction 1 and 2 was defined as a decrease from induction study baseline total Mayo score of ≥3 points and ≥30%, plus a decrease in rectal bleeding subscore of ≥1 point or an absolute rectal bleeding subscore of 0 or 1. [§]Study A3921139 (OCTAVE Open) was ongoing at the time of this interim analysis. [¶]Remission was defined as a total Mayo score of ≤2 with no individual subscore >1, and a rectal bleeding subscore of 0. Adapted from Winthrop KL, et al. *Inflamm Bowel Dis* 2018;24:2258-2265 (in accordance with the CC BY-NC licence)

from OCTAVE Sustain due to treatment failure, as defined in the Supporting Information, and for final efficacy assessments at Week 8 (OCTAVE Induction 1 and 2) or Week 52 (OCTAVE Sustain). In OCTAVE Open, central endoscopic reading was used to determine treatment assignment (based on remission status at baseline) and continuation in the study for induction non-responders at Month 2 (based on whether or not a patient had a clinical response).

2.3 | Statistical analysis

Agreement between central and local reader scores at (a) eligibility screening, (b) Week 8 of OCTAVE Induction 1 and 2, (c) Week 52 of OCTAVE Sustain and (d) Month 2 of OCTAVE Open (induction non-responders) was quantified using weighted kappa statistics.[11] The strength of agreement was interpreted according to the criteria established by Landis and Koch[11] as "slight" (0.00-0.20), "fair" (0.21-0.40), "moderate" (0.41-0.60), "substantial" (0.61-0.80) or "almost perfect" (0.81-1.00).

To assess whether differences were present between the two scoring methods, the agreement between central and local reads was displayed in a four-by-four table based upon the MES categories (Supporting Information), and the kappa statistic was used to evaluate the extent of agreement. Bowker's test, which evaluates the symmetry of the distribution of agreement within a matrix, was used to assess whether or not observed differences in agreement distribution occurred by chance. Analyses were based upon observed data, with no imputation for missing values.

Differences between centrally and locally read MES were assessed using a two-level response (no difference between central and local read; central read ≥1 point higher or lower than local read) or a three-level response (central read ≥1 point lower than local read; no difference between central and local read; central read ≥1 point higher than local read). Potential determinants of disparity between central and local reads (two-level response) at Week 8 of OCTAVE Induction 1 and 2 were assessed using logistic regression analyses. The factors included in these analyses (at induction study baseline, except indicated otherwise) were as follows: age, sex, race, body mass index, prior TNF antagonist failure, prior immunosuppressant failure, oral corticosteroid use, oral corticosteroid dose, extent of disease, disease duration, geographical region (North America vs: Asia; Australia and New Zealand; Eastern Europe; Western Europe; or other), number of patients randomised at site based on induction data (<5 vs ≥5 and <10 vs ≥10), total Mayo score, partial Mayo score, C-reactive protein (CRP) concentration at baseline, CRP concentration at Week 8 and partial Mayo score at Week 8.

In the multivariable logistic modelling process, candidate determinants were evaluated as independent predictors and were selected using a stepwise procedure at a stay criterion and entry criterion of 0.05, with disagreement (two-level) between central and local scoring as the dependent variable. Odds ratios with 95% confidence intervals (CIs) are reported for each factor, representing the effect of influence from the evaluated factor on the disagreement between central and local reads.

Efficacy data, including endoscopic improvement, remission and endoscopic remission based on central and local reads of MES, were analysed for OCTAVE Induction 1 and 2 (pooled) and OCTAVE Sustain. Non-responder imputation was used for missing data, with 95% CIs, based on the normal approximation for the difference in binomial proportions, and P-values to assess the treatment effect based on the Cochran-Mantel-Haenszel chi-squared test (Supporting Information). Point estimates and 95% CIs were calculated for differences between treatments for remission, endoscopic improvement and endoscopic remission estimates, based upon central and local reads at Weeks 8 and 52.

2.4 | Role of the funding source

These studies were funded by Pfizer Inc. The funder of the study had a role in study design, data collection, data analysis, data interpretation and writing of the report. The medical writing support was funded by Pfizer Inc. All authors reviewed and approved the final manuscript, had access to the study data, and accept responsibility for the decision to submit for publication.

3 | RESULTS

3.1 | Patients

Patient demographics and baseline disease characteristics were generally similar across treatment groups and OCTAVE studies, except that lower total Mayo scores, partial Mayo scores and CRP concentrations were observed for participants in OCTAVE Sustain, as expected in this responder population (Table 1).

3.2 | Agreement between central and local endoscopic reads

There was substantial agreement between central and local endoscopic reads at screening (1126/1461 patients [77.1%]; kappa statistic 0.62 [95% CI 0.59-0.66]) and Week 8 (677/1061 patients [63.8%]; kappa statistic 0.62 [95% CI 0.59-0.66]) of OCTAVE Induction 1 and 2, and moderate agreement at Week 52 of OCTAVE Sustain (185/333 patients [55.6%]; kappa statistic 0.56 [95% CI 0.50-0.62]) and for induction non-responders at Month 2 of OCTAVE Open (229/382 patients [59.9%]; kappa statistic 0.54 [95% CI 0.48-0.60]) (Figure 2). The 1461 patients in the screening analysis included all patients who were screened that had both central and local reads; those who only had a read from one method were excluded. When disagreement was present between the methods (22.9%, 36.2%, 44.4% and 40.1% at screening, OCTAVE Induction 1 and 2 Week 8, OCTAVE Sustain Week 52 and OCTAVE Open Month 2 respectively), it was most frequently a discrepancy of 1 point (21.8%, 32.3%, 40.2% and 35.6% respectively), and was predominantly in patients with centrally read scores of 2-3;

discrepancies of 2 or 3 points were uncommon (<5% of patients across all studies and time points).

At screening of OCTAVE Induction 1 and 2, the proportion of patients with a central read higher than the local read (178/1461 patients [12.2%]) was similar to the proportion of patients with a local read higher than the central read (157/1461 patients [10.7%]); statistical testing of the distribution of disagreement showed no significant evidence of asymmetry at screening of OCTAVE Induction 1 and 2 (Bowker's test P = 0.0852). In contrast, statistical testing of the symmetry of the distribution of disagreement showed that the skew in distribution observed towards lower reads by local readers at Week 8 of OCTAVE Induction 1 and 2, Week 52 of OCTAVE Sustain and among induction non-responders at Month 2 of OCTAVE Open was significant (Bowker's test all P < 0.0001). At Week 8 of OCTAVE Induction 1 and 2, the proportion of patients with a central read higher than the local read (287/1061 patients [27.0%]) was substantially higher than the proportion of patients with a local read higher than the central read (97/1061 patients [9.1%]). Similar findings were seen at Week 52 of OCTAVE Sustain (where 113/333 patients [33.9%] had a central read higher than the local read, and 35/333 [10.5%] had a local read higher than the central read) and for induction non-responders at Month 2 of OCTAVE Open (where 126/382 patients [33.0%] had a central read higher than the local read, and 27/382 [7.1%] had a local read higher than the central read) (Figure 2).

Although higher rates of disagreement in local endoscopic reads from central reads were observed for MES 0 and 1 (41.7% [10/24] and 48.6% [34/70]) compared with MES 2 and 3 (24.6% [137/556] and 19.0% [154/811]) at screening of OCTAVE Induction 1 and 2 (Figure 2A), there was no consistent trend of disagreement rates across MES scores in OCTAVE Induction 1 and 2, OCTAVE Sustain and OCTAVE Open (Figure 2).

3.3 | Factors associated with the disparity between central and local reads

Univariate and multivariable logistic regression analyses, to assess potential predictors of disparity between central and local reads, were conducted for Week 8 of OCTAVE Induction 1 and 2. In the univariate analyses, race, geographical region, prior TNF antagonist failure status, baseline total Mayo score, partial Mayo score at baseline and Week 8, and CRP concentration at baseline and Week 8, all had a significant (P < 0.05) association with the disparity between central and local reads at Week 8 of OCTAVE Induction 1 and 2 (Table 2). In the multivariable analysis, lower CRP concentration at Week 8, lower partial Mayo score at Week 8 and not having prior TNF antagonist failure were associated with higher odds of disparity; geographical region was also associated with disparity (Table 2).

At Week 8 of OCTAVE Induction 1 and 2, the proportion of patients with no difference between central and local reads was higher than the proportion with a ≥1-point difference (higher or lower) among all subgroups comprising >10 patients (Supporting Information). When a difference was present, the central read

TABLE 1 Patient demographics and baseline disease characteristics in OCTAVE Induction 1 and 2, OCTAVE Sustain and among induction non-responders in OCTAVE Open

	OCTAVE Induction 1 and 2		OCTAVE Sustain			OCTAVE Open induction non-responders
	Placebo (N = 234)	Tofacitinib 10 mg b.d. (N = 905)	Placebo (N = 198)	Tofacitinib 5 mg b.d. (N = 198)	Tofacitinib 10 mg b.d. (N = 197)	Tofacitinib 10 mg b.d. (N = 429)
Age (years), mean (SD)[a]	41.1 (14.4)	41.2 (13.8)	43.4 (14.0)	41.9 (13.7)	42.9 (14.4)	39.5 (13.6)
Female, n (%)[b]	102 (43.6)	369 (40.8)	82 (41.4)	95 (48.0)	87 (44.2)	168 (39.2)
Geographical region, n (%)[b]						
Europe	135 (57.7)	534 (59.0)	112 (56.6)	113 (57.1)	121 (61.4)	247 (57.6)
North America	53 (22.6)	187 (20.7)	45 (22.7)	39 (19.7)	44 (22.3)	94 (21.9)
Other	46 (19.7)	184 (20.3)	41 (20.7)	46 (23.2)	32 (16.2)	88 (20.5)
Disease duration (years), mean (SD)[c]	8.1 (7.0)	8.1 (7.0)	8.8 (7.5)	8.3 (7.2)	8.6 (7.0)	7.6 (6.5)
Extent of disease, n (%)[b,d]						
Proctosigmoiditis/proctitis[e]	35 (15.0)	133 (14.7)	21 (10.6)	28 (14.3)	33 (16.8)	64 (14.9)
Left-sided colitis	76 (32.6)	307 (34.0)	68 (34.3)	66 (33.7)	60 (30.6)	150 (35.0)
Extensive colitis or pancolitis	122 (52.4)	463 (51.3)	108 (54.5)	102 (52.0)	103 (52.6)	215 (50.1)
Oral corticosteroid use at baseline, n (%)[b]	113 (48.3)	412 (45.5)	105 (53.0)	101 (51.0)	92 (46.7)	179 (41.7)
Prior TNF antagonist failure, n (%)[b]	124 (53.0)	465 (51.4)	89 (44.9)	83 (41.9)	93 (47.2)	261 (60.8)
Total Mayo score, mean (SD)[c]	9.0 (1.5)	9.0 (1.4)	3.3 (1.8)	3.3 (1.8)	3.4 (1.8)	8.6 (1.6)
Partial Mayo score, mean (SD)[c]	6.4 (1.2)	6.4 (1.2)	1.8 (1.4)	1.8 (1.3)	1.8 (1.3)	5.8 (1.4)
MES, mean (SD)[c,f]	2.6 (0.5)	2.6 (0.5)	1.5 (0.9)	1.5 (0.9)	1.6 (0.9)	2.8 (0.5)
CRP (mg/L), median (range)[b,d]	4.7 (0.1-205.1)	4.6 (0.1-208.4)	1.0 (0.1-45.0)	0.69 (0.1-33.7)	0.89 (0.1-74.3)	4.4 (0.1-101.0)

Abbreviations: b.d., twice daily; CRP, C-reactive protein; MES, Mayo endoscopic subscore; N, number of patients in the treatment group; n, number of unique patients with a particular characteristic; SD, standard deviation; TNF, tumour necrosis factor.

[a]Based on data from screening of induction studies for OCTAVE Induction 1 and 2 and OCTAVE Sustain; based on baseline of OCTAVE Open for induction non-responders at Month 2 of OCTAVE Open.

[b]Based on data from baseline of induction studies.

[c]Based on data from baseline of OCTAVE Induction 1 and 2, OCTAVE Sustain or OCTAVE Open.

[d]Based on patients with non-missing values.

[e]One patient with proctitis was enrolled into OCTAVE Induction 2 as a protocol deviation and assigned to receive tofacitinib 10 mg b.d. in OCTAVE Induction 2 followed by tofacitinib 10 mg b.d. in OCTAVE Open.

[f]MES as determined by central read.

was higher than the local read for most patients in all subgroups (Supporting Information). Patients in Asia were the most likely to have disparity (47.8% had a central read ≥1 point higher or lower than the local read), whereas patients in Eastern Europe were the least likely to have disparity (30.9% had a central read ≥1 point higher or lower than the local read). Patients aged ≥50 years were more likely to have a difference between central and local reads than younger age groups (40.7% vs 33.8%-35.3%, respectively). The proportion of patients with disparity was numerically higher among patients with a partial Mayo score <6 at Week 8 vs those with a partial Mayo score ≥6 at Week 8 (40.8% vs 23.8% respectively); the same trend was observed for baseline partial Mayo score (42.7% vs 34.1%). Patients without prior TNF antagonist

failure were more likely to have disparity than those with prior TNF antagonist failure (40.6% vs 32.1% respectively). Patients with a CRP concentration <3 mg/L at induction study baseline were more likely to have disparity than those with a CRP concentration ≥3 mg/L (42.5% vs 32.9% respectively), and the same trend was observed for CRP concentration at Week 8 (CRP <3 mg/L, 42.7%; CRP ≥3 mg/L, 26.9%).

The proportion of patients with no difference between central and local reads was also generally higher than the proportion with a ≥1-point difference (higher or lower) across subgroups in either OCTAVE Sustain or among induction non-responders at Month 2 of OCTAVE Open (Supporting Information). Of note, the descriptive differences between central and local reads by subgroup seen at
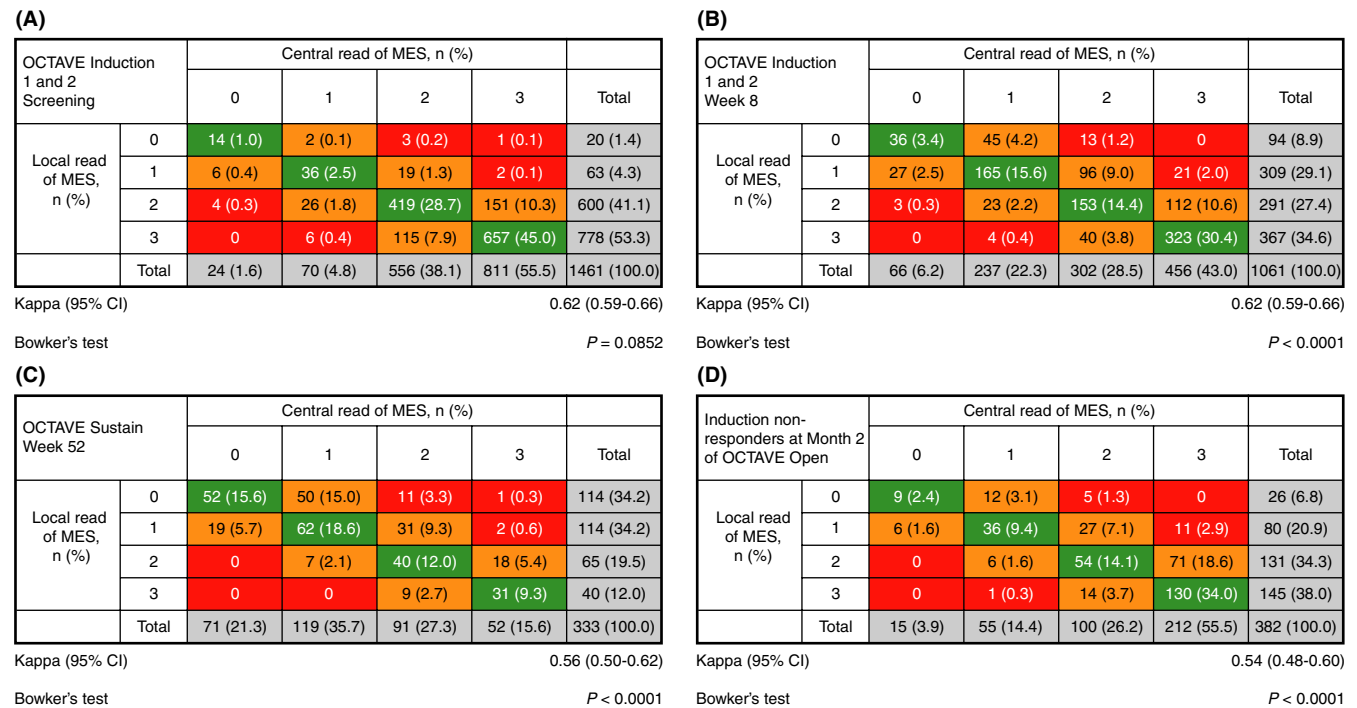
(A)

OCTAVE Induction 1 and 2 Screening		Central read of MES, n (%)				
		0	1	2	3	Total
Local read of MES, n (%)	0	14 (1.0)	2 (0.1)	3 (0.2)	1 (0.1)	20 (1.4)
	1	6 (0.4)	36 (2.5)	19 (1.3)	2 (0.1)	63 (4.3)
	2	4 (0.3)	26 (1.8)	419 (28.7)	151 (10.3)	600 (41.1)
	3	0	6 (0.4)	115 (7.9)	657 (45.0)	778 (53.3)
	Total	24 (1.6)	70 (4.8)	556 (38.1)	811 (55.5)	1461 (100.0)

Kappa (95% CI)   0.62 (0.59-0.66)

Bowker's test   P = 0.0852

(B)

OCTAVE Induction 1 and 2 Week 8		Central read of MES, n (%)				
		0	1	2	3	Total
Local read of MES, n (%)	0	36 (3.4)	45 (4.2)	13 (1.2)	0	94 (8.9)
	1	27 (2.5)	165 (15.6)	96 (9.0)	21 (2.0)	309 (29.1)
	2	3 (0.3)	23 (2.2)	153 (14.4)	112 (10.6)	291 (27.4)
	3	0	4 (0.4)	40 (3.8)	323 (30.4)	367 (34.6)
	Total	66 (6.2)	237 (22.3)	302 (28.5)	456 (43.0)	1061 (100.0)

Kappa (95% CI)   0.62 (0.59-0.66)

Bowker's test   P < 0.0001

(C)

OCTAVE Sustain Week 52		Central read of MES, n (%)				
		0	1	2	3	Total
Local read of MES, n (%)	0	52 (15.6)	50 (15.0)	11 (3.3)	1 (0.3)	114 (34.2)
	1	19 (5.7)	62 (18.6)	31 (9.3)	2 (0.6)	114 (34.2)
	2	0	7 (2.1)	40 (12.0)	18 (5.4)	65 (19.5)
	3	0	0	9 (2.7)	31 (9.3)	40 (12.0)
	Total	71 (21.3)	119 (35.7)	91 (27.3)	52 (15.6)	333 (100.0)

Kappa (95% CI)   0.56 (0.50-0.62)

Bowker's test   P < 0.0001

(D)

Induction non-responders at Month 2 of OCTAVE Open		Central read of MES, n (%)				
		0	1	2	3	Total
Local read of MES, n (%)	0	9 (2.4)	12 (3.1)	5 (1.3)	0	26 (6.8)
	1	6 (1.6)	36 (9.4)	27 (7.1)	11 (2.9)	80 (20.9)
	2	0	6 (1.6)	54 (14.1)	71 (18.6)	131 (34.3)
	3	0	1 (0.3)	14 (3.7)	130 (34.0)	145 (38.0)
	Total	15 (3.9)	55 (14.4)	100 (26.2)	212 (55.5)	382 (100.0)

Kappa (95% CI)   0.54 (0.48-0.60)

Bowker's test   P < 0.0001

FIGURE 2 Distribution of local and central endoscopic reads and weighted kappa statistics at (A) screening in OCTAVE Induction 1 and 2, (B) Week 8 in OCTAVE Induction 1 and 2, (C) Week 52 of OCTAVE Sustain and (D) induction non-responders at Month 2 of OCTAVE Open. CI, confidence interval; MES, Mayo endoscopic subscore; n, number of patients in each category. Data are full analysis set (observed cases), n (%). Agreement between central and local MES assignment: green, no difference; orange, 1-point discrepancy; red, ≥2-point discrepancy. The green boxes, therefore, show a "line" of agreement that runs diagonally across the figure. If a random error was responsible for all of the agreement, discordant scores would be distributed randomly around this "line of agreement." However, if bias were present, discordant scores would be distributed unevenly to one side of the line or the other. The P-values were based upon Bowker's test, which evaluates the distribution (cells that lay off the diagonal line of agreement) of disagreement within the agreement matrix. A P < 0.05 denotes significant asymmetry for exploratory purposes

Week 8 of OCTAVE Induction 1 and 2 were not consistently seen at the other time points (Supporting Information).

3.4 | Efficacy estimates determined by local and central reads

At Week 8 in OCTAVE Induction 1 and 2, and at Week 52 in OCTAVE Sustain, a significantly higher proportion of patients assigned to tofacitinib achieved endoscopic improvement, remission and endoscopic remission relative to placebo, as assessed by both central and local endoscopic reads (Figure 3). In general, the observed rates of endoscopic improvement, remission and endoscopic remission among patients receiving either placebo or tofacitinib were numerically lower for estimates based upon central reads than those derived from local reads (Figure 3). Furthermore, the estimated treatment effect of tofacitinib vs placebo was consistently greater based upon local reads.

4 | DISCUSSION

Centralised reading of endoscopy is widely accepted as a way to minimise bias and decrease measurement variability,[12] a view supported by findings from an induction study that demonstrated "upcoding" of local reads and overestimation of disease activity relative to central reads for determination of trial eligibility.[3] Consequently, a high proportion of patients were enrolled with low disease activity, which increased the placebo rate and reduced the statistical power of the trial to detect a treatment effect.[3]

To our knowledge, our study is the only evaluation of differences between central and local reading performed since the original publication.[3] Unlike the previous study, we did not demonstrate significant "upcoding" during the OCTAVE Induction 1 and 2 screening process. High levels of agreement (77.1%; kappa statistic 0.62 [95% CI 0.59-0.66]) were observed between local and central reads, with no evidence for a systematic difference between the methods. Of note, in the previous study that demonstrated significant "upcoding" at trial baseline, local reads were used to determine eligibility and to generate data for the primary intent-to-treat analysis; central reading was only performed post hoc.[3] In contrast, the OCTAVE Induction trial protocols specified both central and local reads as required procedures at baseline, with eligibility based upon central reads. Since a MES ≥2 was required for enrolment into OCTAVE Induction 1 and 2, most patients screened had a MES of 2 or 3 (by either local or central read); this grouping of MES at the high end of the scale may have contributed to the high levels of agreement seen at screening.

TABLE 2 Univariate and multivariable logistic regression analyses for difference between centrally and locally read endoscopic subscores (two-level response) at Week 8 in OCTAVE Induction 1 and 2

	Univariate logistic regression[a]		Overall P-value	Multivariable logistic regression[b]		Overall P-value
	OR (95% CI)[c]	P-value		OR (95% CI)[c]	P-value	
Age at induction study baseline						
<30 years vs ≥50 years	0.76 (0.54-1.08)	0.1263	0.2771			
30 to <40 years vs ≥50 years	0.74 (0.53-1.04)	0.0818				
40 to <50 years vs ≥50 years	0.79 (0.55-1.13)	0.1941				
Sex						
Female vs male	0.97 (0.75-1.25)	0.8305	0.8305			
Body mass index						
<25 kg/m^2 vs ≥30 kg/m^2	1.16 (0.78-1.71)	0.4640	0.0949			
25 to <30 kg/m^2 vs ≥30 kg/m^2	1.50 (0.98-2.29)	0.0593				
Race						
Black vs white	3.42 (0.81-14.46)	0.0948	0.0093			
Asian vs white	1.64 (1.13-2.37)	0.0085				
Other vs white	1.76 (0.92-3.38)	0.0897				
Geographical region						
Asia[d] vs North America[e]	1.37 (0.87-2.16)	0.1731	0.0065	1.11 (0.69-1.77)	0.6726	0.0025
Australia and New Zealand vs North America[e]	1.07 (0.61-1.88)	0.8230		0.96 (0.53-1.74)	0.8984	
Eastern Europe[f] vs North America[e]	0.67 (0.46-0.96)	0.0284		0.50 (0.34-0.74)	0.0006	
Western Europe[g] vs North America[e]	0.70 (0.49-1.01)	0.0553		0.76 (0.52-1.09)	0.1386	
Other vs North America[e]	1.21 (0.60-2.47)	0.5926		0.95 (0.45-2.00)	0.8964	
Disease duration at induction study baseline						
<6 years vs ≥6 years	0.94 (0.73-1.21)	0.6564	0.6564			
Extent of disease						
Proctosigmoiditis/proctitis[h] vs extensive colitis/pancolitis	1.20 (0.83-1.74)	0.3372	0.1925			
Left-sided colitis vs extensive colitis/pancolitis	1.28 (0.97-1.69)	0.0785				
Oral corticosteroid use at induction study baseline						
No vs yes	0.85 (0.66-1.10)	0.2147	0.2147			
Oral corticosteroid dose at induction study baseline						
<15 mg/day vs none	1.10 (0.75-1.60)	0.6339	0.2462			
≥15 mg/day vs none	1.28 (0.96-1.71)	0.0887				
Other vs none	0.74 (0.37-1.48)	0.3998				
Prior TNF antagonist failure						
No vs yes	1.45 (1.13-1.86)	0.0039	0.0039	1.47 (1.10-1.97)	0.0100	0.0100
Prior immunosuppressant failure						
No vs yes	1.16 (0.88-1.54)	0.2938	0.2938			
Total Mayo score at induction study baseline						
<9 vs ≥9	1.42 (1.10-1.84)	0.0079	0.0079			
Partial Mayo score at induction study baseline						
<6 vs ≥6	1.44 (1.08-1.93)	0.0124	0.0124			
Partial Mayo score at Week 8						
<6 vs ≥6	2.18 (1.60-2.99)	<0.0001	<0.0001	1.88 (1.35-2.61)	0.0002	0.0002
CRP concentration at induction study baseline						
<3 mg/L vs ≥3 mg/L	1.50 (1.16-1.95)	0.0022	0.0022			

(Continues)

TABLE 2 (Continued)

	Univariate logistic regression[a]		Overall P-value	Multivariable logistic regression[b]		Overall P-value
	OR (95% CI)[c]	P-value		OR (95% CI)[c]	P-value	
CRP concentration at Week 8						
<3 mg/L vs ≥3 mg/L	2.00 (1.52-2.62)	<0.0001	<0.0001	1.67 (1.26-2.22)	0.0004	0.0004
Number of patients randomised at site based on induction data						
<5 vs ≥5	1.05 (0.80-1.39)	0.7255	0.7255			
<10 vs ≥10	1.09 (0.84-1.42)	0.5127	0.5127			

Note: Logistic regression analyses were based on a two-level response: no difference between central and local read; central read ≥1 point higher or lower than local read.

Abbreviations: b.d., twice daily; CI, confidence interval; CRP, C-reactive protein; OR, odds ratio; TNF, tumour necrosis factor.

[a]The univariate logistic regression analysis is produced for each factor with treatment group in the model.

[b]A stepwise procedure was used to select factors from the baseline parameters. Factors included in these analyses were (at the induction study baseline, except indicated otherwise): age, sex, race, body mass index, prior TNF antagonist failure, prior immunosuppressant failure, oral corticosteroid use, oral corticosteroid dose, extent of disease, disease duration, geographical region (North America vs: Asia; Australia and New Zealand; Eastern Europe; Western Europe; or other), number of patients randomised at site based on induction data (<5 vs ≥5 and <10 vs ≥10), total Mayo score, partial Mayo score, CRP concentration at induction study baseline, CRP concentration at Week 8 and partial Mayo score at Week 8. The final model included all selected covariates after the selection procedure at the 0.05 level of significance for entry and to stay in the model, which were geographical region, prior TNF antagonist failure, partial Mayo score at Week 8 and CRP concentration at Week 8.

[c]An OR <1 indicates that there were lower odds of disparity (regardless of the direction of the disparity) between central and local reads in the specified subgroup than in the reference subgroup; an OR >1 indicates that there were greater odds of disparity between central and local reads in the specified subgroup than in the reference subgroup.

[d]Japan, Korea and Taiwan.

[e]Canada and the USA.

[f]Croatia, Czechia, Estonia, Hungary, Latvia, Poland, Romania, Russia, Serbia, Slovakia and Ukraine.

[g]Austria, Belgium, Denmark, France, Germany, Israel, Italy, Netherlands, Spain and the UK.

[h]One patient with proctitis was enrolled into OCTAVE Induction 2 as a protocol deviation and assigned to receive tofacitinib 10 mg b.d.

In contrast to the screening results, despite agreement between reading methods in most patients (>50%), systematic differences were found between local and central reads at the other time points. At the end of induction/initiation of OCTAVE Sustain, 63.8% (kappa statistic 0.62 [95% CI 0.59-0.66]) agreement between central and local reads was observed; where there was disagreement, local reader scores were more likely to be lower than those generated by central readers (Bowker's test $P < 0.0001$). Similar results were observed at the end of OCTAVE Sustain/initiation of OCTAVE Open (55.6% agreement; kappa statistic 0.56 [95% CI 0.50-0.62]; Bowker's test $P < 0.0001$) and among induction non-responders at Month 2 of OCTAVE Open (59.9% agreement; kappa statistic 0.54 [95% CI 0.48-0.60]; Bowker's test $P < 0.0001$), when achieving clinical response was necessary for continued eligibility and access to tofacitinib treatment.

One explanation for the difference observed is that both patients and investigators had motivation to continue participation in the maintenance and open-label components of the study. Patients had responded by symptom-based criteria, and a perceived treatment benefit was therefore evident. Local readers, in contrast to central readers, were aware of patients' symptoms, and this may have influenced their endoscopic evaluations. Furthermore, local readers were aware of visit chronology, and may have been influenced by the expectation that patients who successfully completed induction treatment must have received tofacitinib. Central readers

were unaware of information that would lead to such an assumption. Many patients in the OCTAVE studies had failed or were intolerant to conventional and biologic therapies, and consequently had limited treatment options available to them, which may have contributed to "down-coding" of endoscopic scores by local readers to meet criteria for continued participation. Central readers were unaware of patients' prior UC treatments.

Intrinsic limitations of the MES could partly explain why different readers may assign different scores for the same patient, but are unlikely to explain why discrepancies might be skewed in a particular direction. These limitations include the lack of validation, the inability to distinguish superficial ulcers from deep ulcers, the inability to distinguish erythema from marked erythema, and the fact that MES only evaluates the most severely affected visualised segment, with no minimal insertion length. The MES is limited by subjectivity and potential operator variability; however, all sites were trained on scoring MES to limit inter-operator variability in local reads.

Logistic regression analysis was performed on data collected at Week 8 of OCTAVE Induction 1 and 2, to evaluate potential causes of disagreement between the methods; this time point included the largest number of patients who were the most heterogeneous in terms of MES range. Patients with less severe symptoms and a lower inflammatory burden, based on partial Mayo score and CRP concentration, were more likely to show discordance between central and local reads than those with more severe symptoms and a
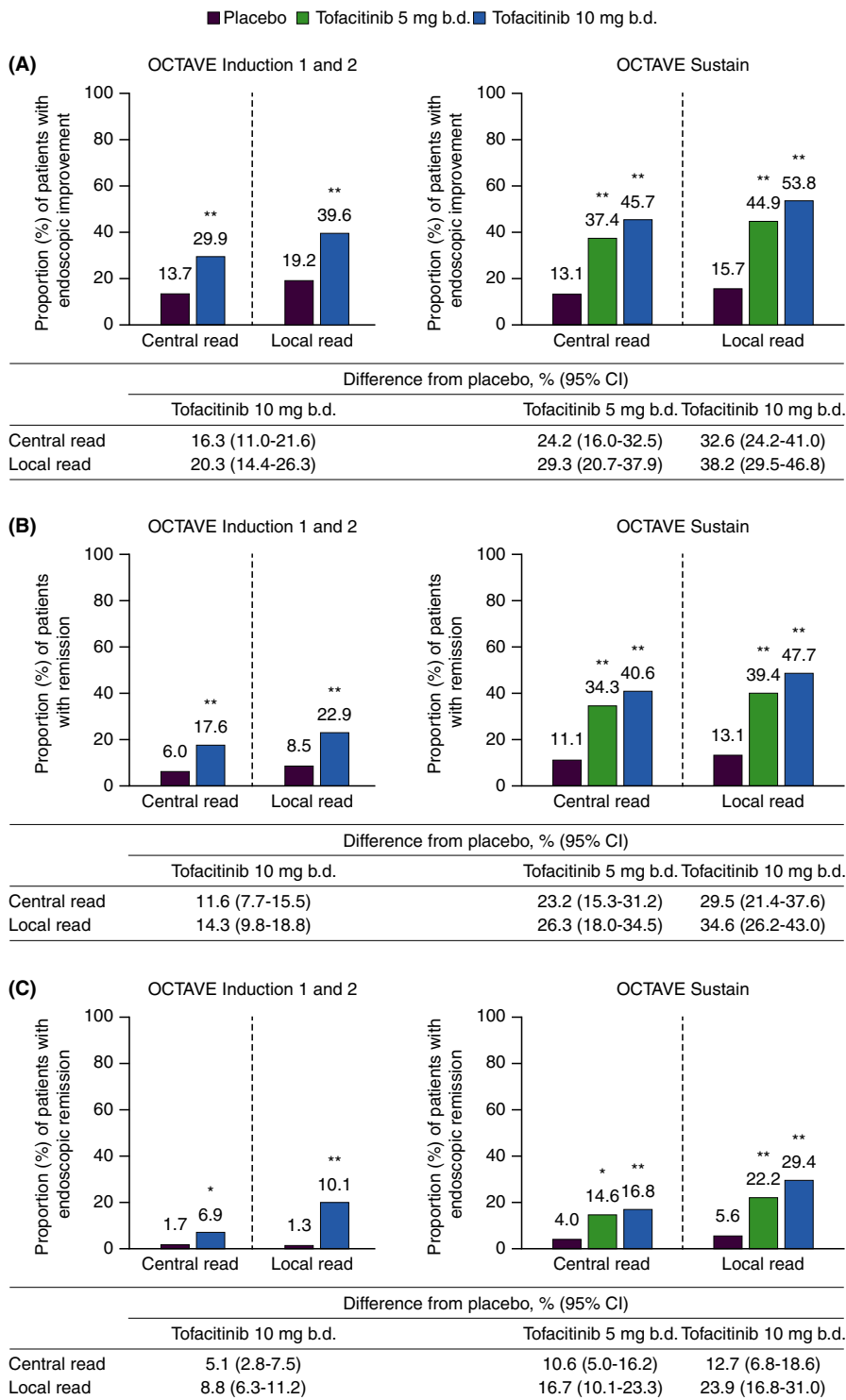
FIGURE 3 (A) Endoscopic improvement,[†] (B) remission[‡] and (C) endoscopic remission[§] at Week 8 (OCTAVE Induction 1 and 2) and Week 52 (OCTAVE Sustain), based on centrally and locally read MES. b.d., twice daily; CI, confidence interval; MES, Mayo endoscopic subscore; TNF, tumour necrosis factor. Data are full analysis set with non-responder imputation; treatment difference from placebo is presented with 95% CIs. [*]$P < 0.01$ vs placebo. [**]$P < 0.001$ vs placebo. For OCTAVE Induction 1 and 2, P-values were based on the Cochran-Mantel-Haenszel chi-squared test, stratified by study, prior TNF antagonist treatment, corticosteroid use at baseline and geographical region. For OCTAVE Sustain, P-values were based on the Cochran-Mantel-Haenszel chi-squared test stratified by treatment assignment in the induction study and remission at baseline. [†]Endoscopic improvement was defined by a MES ≤1. [‡]Remission (primary endpoint) was defined as a total Mayo score ≤2 with no subscore >1, and a rectal bleeding subscore of 0. [§]Endoscopic remission was defined as a MES of 0

higher inflammatory burden. This may reflect a tendency for prior knowledge of a patient's clinical characteristics to bias MES, due to a perception among local readers that endoscopic severity should align with the severity of symptoms/non-endoscopic indicators of severity. Patients without prior TNF antagonist failure were also significantly more likely to have disparity, driven primarily by local readers assigning lower scores than central readers; this could be reflective of TNF antagonist-naïve patients having a less extensive disease or less objective mucosal damage, or the fact that local

readers' scores may have been influenced by an expectation of treatment response.

Patients in Asia were the most likely to have disparity compared with patients in other regions. However, there was no significant ($P > 0.05$) difference between patients located in Asia vs North America; this may have been due to the relatively small number of patients in Asia, or due to disparity being relatively high in North America. Conversely, patients located in Eastern Europe were significantly less likely to have disparity than those in North America

($P < 0.05$). One potential explanation for these differences is cultural variation in how much patients complain of symptoms, which may affect local readers' scores.

Importantly, in patients with moderately to severely active UC in OCTAVE Induction 1 or 2, or OCTAVE Sustain, both central and local reads demonstrated significant efficacy of tofacitinib vs placebo for both induction and maintenance therapy, although treatment effects based on local endoscopic readings were generally numerically greater than those based on central readings.

These findings are subject to some limitations. As this was a post hoc analysis, caution should be applied when interpreting the results; for example differences between central and local reads by subgroup were not always consistent among time points (eg for age and gender), although the reasons for this are unclear. Using a single-read method for central reading may have resulted in more variation among central readers than if multiple reads had been performed. However, previous assessments have found "almost perfect" agreement among central readers with no knowledge of the timing of the endoscopy in relation to the study intervention.[3] Whilst logistic regression analyses evaluated differences between centrally and locally read endoscopic subscores, they did not evaluate which endoscopic features, such as erythema or friability, led to disparity; such information is beyond the scope of this analysis. Finally, this analysis includes data for one agent from a single trial programme and may not be generalisable to other trials.

In summary, although there was agreement between local and central scores for the majority of patients, there was evidence of variability between central and local reads of MES in the OCTAVE clinical programme. Importantly, local reads were systematically lower than central reads, suggesting the possibility that assessments may be affected by bias. Although some potential influencing factors of disparity were identified, further research is required to understand how they may cause disparity between central and local reads. Finally, tofacitinib demonstrated efficacy vs placebo for both induction and maintenance therapy, irrespective of whether central or local reading was used.

ACKNOWLEDGEMENTS

AUTHORSHIP

STUDY ETHICS AND PATIENT CONSENT

All studies were conducted in compliance with the Declaration of Helsinki and the International Conference on Harmonisation Good Clinical Practice Guidelines, and were approved by the Institutional Review Boards and/or Independent Ethics Committees at each of the investigational centres participating in the studies or at a central Institutional Review Board. All patients provided written informed consent.

DATA AVAILABILITY STATEMENT

Upon request, and subject to certain criteria, conditions and exceptions (see https://www.pfizer.com/science/clinical-trials/trial-data-and-results for more information), Pfizer will provide access to individual de-identified participant data from Pfizer-sponsored global interventional clinical studies conducted for medicines, vaccines and medical devices (a) for indications that have been approved in the US and/or EU or (b) in programmes that have been terminated (ie development for all indications has been discontinued). Pfizer will also consider requests for the protocol, data dictionary and statistical analysis plan. Data may be requested from Pfizer trials 24 months after study completion. The de-identified participant data will be made available to researchers whose proposals meet the research criteria and other conditions, and for which an exception does not apply, via a secure portal. To gain access, data requestors must enter into a data access agreement with Pfizer.

ORCID

Brian G. Feagan [iD] https://orcid.org/0000-0002-6914-3822
Walter Reinisch [iD] https://orcid.org/0000-0002-2088-091X

REFERENCES

1. Vuitton L, Peyrin-Biroulet L, Colombel JF, et al. Defining endoscopic response and remission in ulcerative colitis clinical trials: an international consensus. *Aliment Pharmacol Ther.* 2017;45:801-813.
2. D'Haens G, Sandborn WJ, Feagan BG, et al. A review of activity indices and efficacy end points for clinical trials of medical therapy in adults with ulcerative colitis. *Gastroenterology.* 2007;132:763-786.
3. Feagan BG, Sandborn WJ, D'Haens G, et al. The role of centralized reading of endoscopy in a randomized controlled trial of mesalamine for ulcerative colitis. *Gastroenterology.* 2013;145:149-157.
4. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER) Ulcerative colitis: clinical trial endpoints. Guidance for industry. 2016. http://www.fda.gov/downloads/Drugs/Guidances/UCM515143.pdf. Accessed July 8, 2021.
5. European Medicines Agency Guideline on the development of new medicinal products for the treatment of ulcerative colitis. 2018. https://www.ema.europa.eu/documents/scientific-guideline/guideline-development-new-medicinal-products-treatment-ulcerative-colitis-revision-1_en.pdf. Accessed July 8, 2021.
6. Sands BE, Sandborn WJ, Panaccione R, et al. Ustekinumab as induction and maintenance therapy for ulcerative colitis. *N Engl J Med.* 2019;381:1201-1214.
7. Sandborn WJ, Su C, Sands BE, et al. Tofacitinib as induction and maintenance therapy for ulcerative colitis. *N Engl J Med.* 2017;376:1723-1736.

8. Vermeire S, O'Byrne S, Keir M, et al. Etrolizumab as induction therapy for ulcerative colitis: a randomised, controlled, phase 2 trial. *Lancet*. 2014;384:309-318.

9. Sandborn WJ, Ghosh S, Panes J, et al. Tofacitinib, an oral Janus kinase inhibitor, in active ulcerative colitis. *N Engl J Med*. 2012;367:616-624.

10. Lichtenstein GR, Loftus EVJ, Soonasra A, et al. Tofacitinib, an oral Janus kinase inhibitor, in the treatment of ulcerative colitis: an interim analysis of an open-label, long-term extension study with up to 5.5 years of treatment [abstract]. *Am J Gastroenterol*. 2019;114:S413-S414. Abstract 704.

11. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159-174.

12. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER) Clinical trial imaging endpoint process standards. Guidance for industry. 2018. https://www.fda.gov/regulatory-information/search-fda-guida nce-documents/clinical-trial-imaging-endpoint-process-standards-guidance-industry. Accessed July 8, 2021.

SUPPORTING INFORMATION

Additional supporting information will be found online in the Supporting Information section.

How to cite this article: Feagan BG, Khanna R, Sandborn WJ, et al. Agreement between local and central reading of endoscopic disease activity in ulcerative colitis: results from the tofacitinib OCTAVE trials. *Aliment Pharmacol Ther*. 2021;54:1442–1453. doi:10.1111/apt.16626