Neurogastroenterology & Motility

WILEY

Neurogastroenterology & Motility NGM

ORIGINAL ARTICLE OPEN ACCESS

Intra- and Interobserver Variability of Acute Food–Induced Reactions During Confocal Laser Endomicroscopy: An International Multicenter Validation Study

Lukas Michaja Balsiger[1,2] | Tom van Gils[3] | Yaser Hatem[4] | Amanda Blomsten[3] | Karlien Raymenants[1] | Cedric Van de Bruaene[1] | Leila Juvyns[1] | Johann P. Hreinsson[3] | Tim Vanuytsel[1] | Hans Törnblom[3] | Christian Sina[4,5] | Magnus Simren[3,6] | Jan Tack[1,3]

[1]Translational Research Center for Gastrointestinal Disorders, KU Leuven, Leuven, Belgium | [2]Department for Biomedical Research, University of Bern, Bern, Switzerland | [3]Department of Molecular and Clinical Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden | [4]University of Luebeck Institute of Nutritional Medicine, University Hospital of Schleswig-Holstein, Campus Lübeck and University of Lübeck, Lübeck, Germany | [5]Fraunhofer Research Institution of Individualised and Cell-Based Medical Engineering (IMTE), Lübeck, Germany | [6]Center for Functional GI and Motility Disorders, University of North Carolina/Chapel Hill, Chapel Hill, North Carolina, USA

Correspondence: Jan Tack (jan.tack@kuleuven.be)

Received: 21 November 2024 | **Revised:** 3 February 2025 | **Accepted:** 13 March 2025

Funding: LMB received funding from a Postdoc Mobility grant from the Swiss National Science foundation (P5R5PM_225320 and P500PM_206612) and from the Novartis Foundation for medical-biological Research (#23C149). KR is funded by a research fellowship of the Flanders Research Foundation (FWO Vlaanderen (1128723N)). TV is supported by the Flanders Research Foundation (FWO) through a senior clinical research mandate (1830517N) and a research grant (G059822N). The University of Lübeck received funding from the Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung (BMBF)). The study interventions conducted in Lübeck were embedded in the large-scale project INDICATE-FH, Funding number: 01EA2109.

Keywords: atypical food allergies | CLE | confocal laser endomicroscopy | IBS | irritable bowel syndrome

ABSTRACT

Background and Study Aims: Probe-based confocal laser endomicroscopy (pCLE) enables real-time microscopic visualization of the duodenal mucosa and has shown acute food–triggered disruption of the duodenal epithelial barrier of patients with irritable bowel syndrome (IBS). The interpretation of the recordings is subjective, with unknown agreement rates. The aim of this study was to investigate the intra- and interobserver variability of this technique.

Patients and Methods: An international multicenter study was performed, including pCLE recordings from three centers. Recordings were randomized and re-evaluated by five blinded experienced assessors. Low-quality recordings were excluded. The mucosa was considered altered if both fluorescein leakage and luminal particles were observed. Agreement was quantified using Fleiss' and Cohen's kappa (κ). Reference videos (i.e., videos with 100% agreement) were used to assess the optimal characteristics of videos needed to make a judgment based on the optimal receiver operating characteristic curve cutoff.

Results: Of the 119 individual recordings, 87 could be used for analyses (total of 86,408 frames). Intraindividual agreement rate was 80%–100%, whereas the interindividual agreement rate was 85% ($\kappa = 0.68$). The agreement rate with the endoscopist ranged 54%–95% ($\kappa = 0.15$–0.89). The optimal cutoff to distinguish altered from unaltered was by observing alterations in ≥ 2 out of 6 mucosal spots (100% sensitivity and specificity).

Conclusion: Our study showed a substantial to perfect intraobserver agreement and a substantial interobserver agreement for the judgment of acute food-triggered disruption of the duodenal epithelial barrier by pCLE, confirming that this real-time readout is reliable and reproducible.

Lukas Michaja Balsiger and Tom van Gils equal contributions.

Summary

- Acute food induced mucosal alterations have been visualized in the duodenum using confocal laser endomicroscopy but the reproducibility of these findings have not been previously assessed.
- Agreement rates between real-time in vivo imaging and blinded post hoc assessment were substantial. Post hoc agreement rates between blinded assessors from different centers was substantial.
- Current criteria for acute food induced mucosal alterations allow reproducible and comparable detection of alterations.

1 | Introduction

Confocal laser endomicroscopy (CLE) is a technique allowing for real-time microscopic visualization of the mucosa during various endoscopic procedures [1]. Assessment of the duodenal mucosa during esophago-gastro-duodenoscopy (EGD) in patients with irritable bowel syndrome (IBS) has shown acute food–Induced mucosal alterations within minutes after intraluminal food administration [2, 3]. This procedure has been termed Food Allergy Sensitivity Testing (FAST) by the pioneers of this application of CLE [4]. The observed alterations included fluorescein leakage, epithelial breaks, and widening of the intervillous space due to fluorescein leaking through the mucosa [2, 3]. Previous studies found these reactions only in patients with IBS but not in disease controls [2, 3]. Furthermore, IBS symptoms improved after excluding the trigger food (i.e., food leading to acute alterations) from the diet of the patients [3]. In contrast, a recent study found no difference in acute alterations following lipid infusion between IBS patients and healthy controls [5]. Based on the most recent white paper, acute mucosal alterations observed on CLE are defined as fluorescein leakage, along with the appearance of particles in the lumen (previously termed "cell shedding") [4]. These criteria are based on the previously reported findings in animal models that impaired mucosal integrity triggered by inflammatory stimuli was characterized by the shedding of cells and leakage of contrast agent [6]. However, the exact pathophysiological mechanisms underlying these observed alterations are unknown.

Additionally, the reliability and reproducibility of these CLE findings have not been systematically assessed. The currently available CLE system is probe-based (pCLE) (Cellvizio, Mauna Kea technologies, Paris). When performing pCLE, the microscopy probe is introduced through the working channel of the endoscope, which differs from the previously used and no longer available system where the microscope was built-in to the endoscope [2, 3]. Testing the reliability and reproducibility of food-induced mucosal alterations visualized using pCLE is especially important since the limited available literature proposes judging pCLE videos in real time during the procedure [7–10]. Furthermore, there is no gold standard available for acute mucosal reactions to compare with, rendering validation of the observed findings difficult. This is in contrast to other diseases where the reliability and reproducibility of pCLE findings have been studied and compared to histopathological findings, such as Barrett's esophagus and pancreatic cyst lesions [11, 12].

Therefore, we aimed to assess the intra- and interobserver variability of food-induced mucosal alterations in the duodenum visualized using pCLE.

2 | Methods

2.1 | Study Design

This is a multicenter, blinded study to assess the reproducibility of pCLE to assess food-induced mucosal alterations. To this end, we evaluated intra- and interobserver variability and the learning curve of this technique. This study was conducted by an international panel of investigators from three European centers involved in ongoing prospective pCLE research: KU Leuven (Belgium), University of Gothenburg/Sahlgrenska University Hospital (Sweden) and University Hospital Schleswig-Holstein, Campus Lübeck (Germany). The videos evaluated during this study originated from one multicenter and two single center studies investigating food-induced mucosal alterations in the duodenum using pCLE in patients with IBS (DRKS00029323, DRKS00013534, NCT05097872, and NCT06413004). Studies were approved by each local institutional review board, and all study participants provided witnessed informed consent before undergoing any study-related procedure (Internal reference number Leuven: S65495, Gothenburg: 2021-04336 and 2022-06370-02, Lübeck: 2022-111 and 2022-547_1).

2.2 | Selection of Assessors

The most experienced assessors from each center participated in the video revisions: two assessors from both Gothenburg and Leuven, and one assessor from Lübeck. Previous experience of the assessors is provided in Table S1. All five assessors completed the CLE course provided by Mauna Kea Technologies via an online platform (https://www.cellvizio.net/learn/image-interpretation/digestive/9-irritable-bowel-syndrome—last accessed on July 10, 2024). All endoscopists performing the exams received training by Mauna Kea, either online or in person, or by experienced peers, prior to performing the exams used in this study.

Two inexperienced assessors (previous experience in Table S1) from one center conducted the same video revision on the center's own videos to assess the learning curve. Both assessors completed the CLE FAST course provided by Mauna Kea Technologies. In the 2 months following the initial video revision, they attended at least 7 more CLE exams in real life and followed a structured teaching intervention (outline provided in Appendix S1) before conducting video revision on a random selection of videos from the initial assessment for the second time.

2.3 | Probe-Based CLE

Upon reaching the duodenum and confirming its normal appearance during standard EGD, fluorescein was applied intravenously as a contrast agent and the microscopy probe was gently

placed on the duodenal surface through the working channel of the endoscope. Evaluation was performed at baseline and after the sequential administration of food. The presence of both intraluminal fluorescein and luminal particles (previously termed "cell shedding") was defined as altered mucosa (previously termed "positive"), and the absence of both intraluminal fluorescein and luminal particles as unaltered mucosa (previously termed "negative") [4, 7, 8]. Figure 1 depicts altered and unaltered mucosa. Several separate mucosal areas (termed "spots") were visualized for approximately 30 s each before moving on to the following spot. Although no universally established cutoff exists, the global judgment of altered mucosa required visible alterations in several separate spots. In contrast, one single altered spot was not enough to consider a globally altered mucosa. This judgment was based on real-time visualization of at least 4 separate mucosal spots by the endoscopist as a binary outcome: altered (previously termed "positive") or non-altered (previously termed "negative") mucosa. After visualization of altered mucosa (i.e., 2/4 or more spots with alterations), the procedure was stopped, and no further food was applied.

While the general protocol is similar in all centers, Table S2 highlights differences in the center-specific protocols.

2.4 | Selection and Re-Assessment of pCLE Videos for the Current Study

Every center compiled videos from nine randomly chosen individuals enrolled in their respective studies. If an individual underwent two pCLE exams, all videos from both exams were included. The only reason for exclusion of a specific individual was if the final judgment of the endoscopist was doubtful.

Videos were extracted in the proprietary .mkt format to avoid data compression losses. Every video represented the recording of one condition in one individual (i.e., 1 video for baseline, 1 video after every food administration). All videos were de-identified and named using two letters per center, followed by a randomly assigned number from 0 to 2000 (e.g., Belgium: BE0022). While blinding to the center was not feasible, investigators were blinded to the order of the video (i.e., baseline or after food administration), the food administered, and the patient characteristics. All videos were recorded at 8 frames/s and played in the Cellvizio viewer—the proprietary software for reviewing .mkt files. First, videos were played in real time without pausing or rewinding—thus reproducing the real-time evaluation during endoscopy. In a second session, all videos were reviewed at half speed (4 frames/s) with the possibility of pausing and rewinding freely.

Assessors quantified the number of different mucosal spots visualized. Different spots were identified by loss of contact with the mucosa (i.e., blackout of the image) and rapid movement suggesting displacement of the probe elsewhere on the mucosa. For every spot, investigators assessed the presence or absence of fluorescein leakage and luminal particles. Since no universal validated cutoff exists on the presence or absence of mucosal alterations, every investigator provided a global assessment of the mucosa on a 6-point Likert scale ranging from 1 = "definitely unaltered" to 6 = "definitely altered" (Figure 2). Video quality was assessed for every recording on a 6-point Likert scale ranging from 1 = "insufficient quality, impossible to judge" to 6 = "optimal quality, judgement possible" according to the subjective judgment of the assessors (Table S3). Table S4 provides an overview of all assessments conducted for every video.

The methodology of post hoc assessment and scoring was extensively discussed in several preceding online meetings between the assessors. There was no exchange of information regarding post hoc assessment before all assessments had been completed and logged in the data file.

To assess intraobserver variability, 5 randomly selected videos were included in duplicates in the dataset. They were saved under different filenames, and assessors were not aware of which videos were duplicated.

2.5 | Objectives

The primary objectives were to assess the interobserver variability between post hoc assessors in real-time as well as interobserver variability between post hoc assessors and endoscopists in real-time. These assessments were done based on the binary outcome "altered" versus "unaltered."

Further objectives were to assess the intraobserver variability for post hoc assessment and the interobserver variabilities between

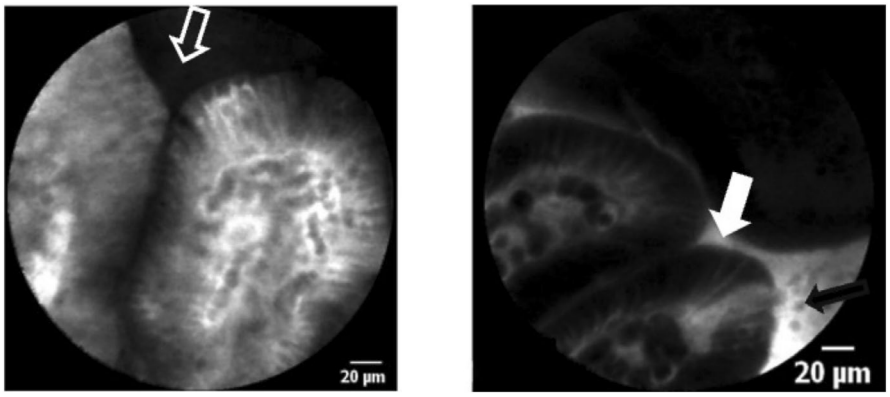

FIGURE 1 | CLE images of the duodenal mucosa. Left panel: Unaltered mucosa with intervillous space devoid of fluorescein (empty arrow). Right panel: Altered mucosa with intervillous appearance of fluorescein (solid white arrow) and particles (solid black arrow) in the intervillous space.

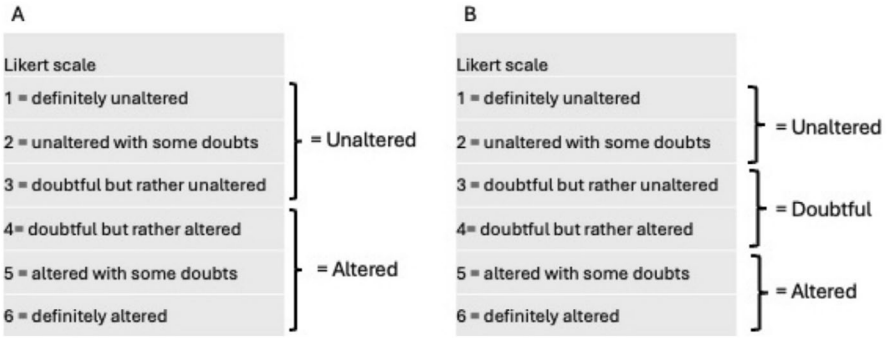

FIGURE 2 | The 6-point Likert scale used to evaluate videos ranging from definitely unaltered to definitely altered. Panel A: Transformation to the binary outcome "altered" versus "unaltered." Panel B: Transformation to a three-category outcome that also included "doubtful."

post hoc assessors in real-time and endoscopists from the same center. Sensitivity analyses were performed with only 4 out of 5 assessors, removing one assessor at a time, to assess whether a single observer induced a strong bias by severely over- or underperforming. Further sensitivity analyses were carried out using a three-point outcome ("unaltered," "doubtful," or "altered").

The interobserver variability of inexperienced assessors with their experienced counterparts and with the endoscopist from the same center was assessed before and after a structured training intervention.

2.6 | Data Transformation

All analyses were conducted after transformation of the Likert scale to a binary outcome ("altered" or "unaltered") as well as a three-point outcome ("altered," "doubtful," or "unaltered"). For the binary outcome, altered mucosa was defined as a rating of 4 or above on the Likert scale (Figure 2A). For the 3-point outcome, ratings 1 and 2 on the Likert scale were transformed to "unaltered," 3 and 4 to "doubtful," and 5 and 6 to "altered" (Figure 2B). Sufficient quality was required for further analysis; therefore, all videos with a median quality judgment of all assessors of ≤ 2 were excluded from further analysis.

2.7 | Statistics

Interobserver variability was assessed using Cohen's kappa (κ) in case of two assessors, Fleiss' κ in the case of three or more assessors [13, 14]. In case the outcome was not binary, ordinally weighted kappa was calculated. Kappa values were interpreted as proposed by Landis and Koch [15]: <0 = poor agreement, 0.01–0.20 = slight agreement, 0.21–0.40 = fair agreement, 0.41–0.60 = moderate agreement, 0.61–0.80 = substantial agreement, 0.81–1.00 = almost perfect agreement. Optimal cutoffs of the receiver operating characteristic (ROC) curve were calculated with the Youden index [16]. Agreement rates are presented as percentages and corresponding kappa values.

All calculations were done in Rstudio (R version 4.3.2 (31-10-2023)—"Eye Holes," Copyright (©) 2023, The R Foundation for Statistical Computing) using the irrCAC package for Fleiss' kappa, irr package for Cohen's kappa, and the pROC package for ROC calculations [17–19].

3 | Results

A total of 119 individual videos were analyzed (total duration of real-time recordings 197 min, total frame numbers 94,572). Within the three centers, the median recording length per video was 229, 669, and 1023 frames, respectively. Overall, a lower frame count was associated with a lower overall quality assessment by the blinded assessors. Consequently, median quality judgment between centers ranged from 2 to 5.

After removing videos with a quality of 2 or less, a total of 87 videos with a total of 86,408 frames were used for further analysis. Median frame count for these videos was 991 (IQR: 547–1139) corresponding to a median duration of 124 s. A flowchart of assessed videos and videos included in the final analyses is presented in Figure S1.

3.1 | Intraobserver Agreement Post Hoc

The intraobserver agreement rate for single assessors ranged from 80% to 100% when using a binary outcome of "altered" versus "unaltered" and judged in real time. Using a three-category outcome, the agreement rate was similar (60%–100%). Slow-motion assessments did not result in substantially different agreement rates. Intraobserver agreement rates for binary and three-category assessments in both real time and slow motion are summarized in Table 1.

3.2 | Interobserver Agreement Post Hoc

Interobserver agreement rate for all videos showed 85% agreement rate between all post hoc assessors ($\kappa = 0.68$). Sensitivity analyses showed no strong effects of a single assessor on the outcome, with agreement rates of 84%–87% ($\kappa = 0.65$–0.73). The agreement rates were not different when post hoc assessors only judged videos from their own center (agreement rates 85%–88%, $\kappa = 0.69$–0.73)—this metric was not available for the center with only one assessor. Including the category "doubtful" using the three-category outcome did not lead to relevant differences in overall interobserver variability (Table 1).

Slow-motion revision did not lead to different interobserver agreement rates, with overall interobserver agreement (all assessors assessing all videos) being 81% ($\kappa = 0.58$).

TABLE 1 | Agreement rates in percent for binary outcome ("altered" vs. "non-altered"), three-category outcome (including "doubtful" judgment), in real-time and slow-motion assessment of food-induced mucosal reactions measured by probe-based confocal laser endomicroscopy (pCLE). Intraobserver agreement rate is represented as a range of individual assessors, agreement rates with endoscopist are presented as range between post hoc assessors with endoscopists from the same center. No three-point agreement rate with endoscopist was assessed as endoscopists' judgments were purely binary.

	Post hoc real-time assessment		Post hoc slow-motion assessment	
	Binary	Three-category	Binary	Three-category
Intraobserver	80%–100%	60%–100%	80%–100%	60%–100%
Interobserver overall	85% (0.68)	68% (0.48)	81% (0.58)	65% (0.42)
Post hoc with the endoscopist within centers	54%–90% (0.24–0.79)	/	46%–88% (0.06–0.64)	/
	% Agreement (kappa)		% Agreement (kappa)	

TABLE 2 | Agreement rates between post hoc assessors and real-time judgment by the endoscopists by center presented as percentage agreement (left panel) and Cohen's kappa (right panel). Within center agreement rates (i.e., post hoc assessor and endoscopist from the same center) are marked in green.

	Center				Center		
Assessor	A	B	C	Assessor	A	B	C
1	82	77	78	1	0.56	0.55	0.4
2	82	77	95	2	0.52	0.55	0.89
3	91	54	88	3	0.75	0.24	0.73
4	85	62	90	4	0.53	0.33	0.79
5	79	54	90	5	0.51	0.15	0.79
	Agreement (%)				Cohen's kappa		

3.3 | Agreement Between Post Hoc and Endoscopist Within and Between Centers

The agreement rate between post hoc assessors and the judgment of the endoscopists within the same center was high for two centers, ranging from 82% ($\kappa = 0.52–0.56$) to 90% ($\kappa = 0.79$) but was substantially lower in the third center, 54% ($\kappa = 0.24$). The agreement rates between post hoc assessment and the judgment of the endoscopists from different centers were similar to the agreement rates within centers (range of agreement 54%–95%, $\kappa = 0.15–0.89$); the detailed distribution is presented in Table 2. Agreement rates within centers were similar in slow motion (Table 1).

Because the judgments of the endoscopists were always binary, the three-point outcome was not assessed.

3.4 | Gold Standard Videos

Of all videos that were included in the analyses, 59 showed agreement rates of 100% between the blinded assessors (18 altered, 45 unaltered). In these videos, the agreement with real-time judgment by the endoscopist was 97% ($\kappa = 0.92$). Due to the full agreement rate of blinded assessors, these videos were further used as reference videos to determine cutoffs to optimize the performance of imaging.

Reference videos showed a mean of 5.5 (95% CI: 2.8–8.0) visualized mucosal spots. In videos judged as unaltered, a mean of 3% of spots were altered. In contrast, the altered videos showed a mean of 61% altered spots. The receiver operating characteristic (ROC) curve analysis identified a cutoff of 31% of altered spots as optimal to make the global judgment of altered versus unaltered (specificity, sensitivity, positive predictive value, and negative predictive value = 100%) (Figure S2). Thus, using these reference videos, the optimal number of mucosal spots to visualize was found to be 6, and a cutoff of 2 or more of these 6 spots showed optimal performance in distinguishing altered from unaltered mucosa. Characteristics of reference videos are summarized in Table 3.

3.5 | Learning Curve

The agreement rate of inexperienced assessors with their more experienced counterparts from the same center varied from 75% to 82% ($\kappa = 0.52–0.63$). Similarly, agreement rates with the endoscopist from the same center ranged from 70% to 82% ($\kappa = 0.45–0.62$). This contrasted with the more experienced counterparts who presented agreement rates between themselves of 85% ($\kappa = 0.69$) and between each post hoc assessor and the endoscopist of 90% ($\kappa = 0.79$).

Following additional exposure and a structured training intervention, the agreement rate between inexperienced and

TABLE 3 | Characteristics of reference videos stratified by videos regarded as "altered" or "non-altered." Data are presented as means and proportions (95% CI). *p* values are obtained using an unpaired *t* test and Fisher's exact test respectively.

	Judged as altered ($n=18$)	Judged as non-altered ($n=45$)	*p*
Imaged spots	6 (5–7)	5 (5–6)	0.068
Proportion of altered spots	61% (54%–68%)	3% (2%–5%)	<0.001

experienced assessors increased to 83%–92% (kappa = 0.65–0.83) as did the agreement rate of inexperienced assessors and the endoscopist (83%–92%; $\kappa = 0.67$–0.83).

4 | Discussion

We report for the first time the intra- and interobserver agreement rates of acute food-induced mucosal reactions assessed by pCLE in patients with IBS. We found high intra- and interobserver agreement rates.

The interobserver agreement rate between assessors was substantial when videos were assessed in real-time speed without the possibility of pausing or rewinding. Importantly, the agreement rate between real-time assessment by the endoscopist and post hoc blinded assessment was also moderate to substantial, except for one center with a lower agreement rate. One contributing factor to this finding might be that several endoscopists conduct pCLE in this center, which is in contrast with the other centers using only one or a small group of experienced operators. This hypothesis is further strengthened by the fact that all post hoc assessors showed relatively low agreement rates with the endoscopists from that center.

The kappa values we report here are similar to kappa values recently reported for radiological findings such as intrahepatic duct dilation in patients being evaluated for suspected cholecystitis as measured by ultrasound (US) and computer tomography (CT) [20]. This is remarkable as CT and US are established imaging modalities for the diagnostic workup of suspected cholecystitis, have been used in daily clinical practice for years, and radiologists have undergone years of dedicated training for image evaluation. We report substantial agreement between experienced investigators for this experimental technique, lending credibility to the robust evaluation of the binary outcome of "altered" versus "unaltered" mucosa. Interestingly, slow-motion assessment did not increase the agreement rates between assessors. One can speculate that several factors influence this finding: first, all assessors were trained in and accustomed to real-time assessment. Furthermore, reading a slow-motion video was reported by all assessors to cause more fatigue—thus potentially leading to increased inattention bias. This was further reinforced by the fact that slow-motion assessment took place after real-time assessment.

Using full agreement as a criterion, we identified reference videos that allowed us to identify characteristics of ideal recordings. Reference videos presented an average of 6 different mucosal spots, 2 or more of which had to be altered (= present fluorescein leakage and luminal particles) to be globally considered altered. Agreement rates between blinded assessors and endoscopists for reference videos were excellent, further reinforcing their validity as ideal recordings. Different approaches have been reported previously, ranging from the imaging of 3 [9, 10] to 5 mucosal spots [2]. Based on our reference videos, we propose that videos should ideally depict at least 6 different mucosal spots, and the mucosa should be considered altered if at least 1/3 of mucosal spots show alterations—however, this requires further assessment and validation.

We demonstrated that the interobserver agreement rates between inexperienced and experienced assessors from the same center and between inexperienced assessors and the same center's endoscopist were lower than in the group of experienced assessors. After following several additional exams and a structured training using reference videos, these assessors showed agreement rates that were comparable to the experienced counterparts from the same center. While the agreement rate increased, it is unclear whether this is relevant as no cutoffs for agreement rates or other quality measures have been established.

Strengths of the study include the use of videos from multiple centers and the rigorous blinding. The centers involved are the only European centers currently using this technique in prospective research protocols to the best of our knowledge, and the assessors that performed post hoc assessment are the most experienced assessors from these respective centers. By evaluating the real-time videos without rewinding or pausing, we reproduced the setting of performing the exam while judging the images in real-time to the best of our technical abilities. Studies on the reliability of new techniques such as pCLE are needed to quantify the performance characteristics and reproducibility of the technique before implementation in further research and potentially clinical use. Our confirmation of this assessments' reliability paves the way for further studies into the underlying pathophysiology and potential clinical implications of food-induced mucosal alterations. More study into the relevance of these alterations is merited, and studies should include healthy volunteers to assess disease specificity, as recent studies have also shown acute alterations in healthy individuals [5].

Limitations of this study include limitations inherent to this exam. Most importantly, there is no gold standard to compare the assessments with. This is in contrast with other uses of this technique—such as Barrett's esophagus–where pCLE images can be compared to globally accepted reference standards such as histology [11]. Furthermore, the criteria of altered mucosa (i.e., fluorescein leakage and "cell shedding") were based on phenomena identified in animal models following intraperitoneal inflammatory stimuli and were assessed using another imaging platform and have themselves never been rigorously validated for this use [6]. Thus, while the agreement rate of these findings was substantial, it remains unclear which pathophysiological mechanisms lead to the observed alterations and whether these alterations carry any clinical relevance. Because no universal cutoffs have been previously determined and to reflect different degrees of certainty, we decided to make the global assessment on a 6-point Likert scale. This was an arbitrary scale and can be criticized as we later dichotomized our 6-point Likert scale to the binary

outcome of "altered" versus "unaltered" to better reflect the dichotomous outcome during real-time endoscopy. Although dichotomization reduced granularity, agreement rates did not differ when including a category of "doubtful" interpretations. The ideal scale to be used remains to be determined. As reactions are not always clear cut, leaving the option to indicate certain level of doubt might help assessors' judgment—this requires further validation to ensure validity. Additional limitations include the absence of endoscopic images during post hoc assessments; assessors were not able to follow the positioning of the probe using an endoscopic image nor were they able to see mucosal lesions or the presence of bile—both of which may lead to altered microscopic images [4].

In conclusion, we found substantial intra- and interobserver agreement rates regarding the presence and absence of acute food-induced alterations of duodenal mucosa in the blinded revision of pCLE images when judged in real time. Our findings support the use of these criteria with regard to reproducibility and comparability, but the clinical and pathophysiological meaning of these findings requires further studies.

Author Contributions

L.M.B., T.G.: Study initiation, conceptualization, and planning of international collaboration, image interpretation post hoc, statistical analyses, manuscript writing; Y.H., A.B., K.R., C.V.B., L.J.: Image interpretation post hoc, conceptualization of post hoc interpretation, manuscript revision; J.P.H.: performed part of the CLE endoscopies in Gothenburg, statistical analyses, manuscript revision; T.V., H.T.: study conceptualization, manuscript revision; C.S., M.S.: study conceptualization, infrastructural support, principal investigators of trials using CLE and generating videos, manuscript revision; J.T.: study conceptualization, infrastructural support, principal investigator for ongoing CLE trials and performed all CLE endoscopies in Leuven, manuscript revision.

Conflicts of Interest

C.V.B. received a speaker fee from Mayoly and the Rome Foundation, H.T. served as Consultant/Advisory Board member for Allergan, Cinclus Pharma, Medifactia, and VIPUN, and as a speaker for Galapagos, Tillotts, and Takeda. All the other authors report no conflicts of interest.

Data Availability Statement

The authors have nothing to report.

References

1. R. Kiesslich, "Diagnostic Value of Endomicroscopy for Gastrointestinal Diseases: New Possibilities and Concepts," *Techniques and Innovations in Gastrointestinal Endoscopy* 23 (2021): 57–68, https://doi.org/10.1016/j.tige.2020.09.005.

2. A. Fritscher-Ravens, D. Schuppan, M. Ellrichmann, et al., "Confocal Endomicroscopy Shows Food-Associated Changes in the Intestinal Mucosa of Patients With Irritable Bowel Syndrome," *Gastroenterology* 147, no. 5 (2014): 1012–1020.e4, https://doi.org/10.1053/j.gastro.2014.07.046.

3. A. Fritscher-Ravens, T. Pflaum, M. Mösinger, et al., "Many Patients With Irritable Bowel Syndrome Have Atypical Food Allergies Not Associated With Immunoglobulin E," *Gastroenterology* 157, no. 1 (2019): 109–118.e5.

4. R. Kiesslich, M. Rusticeanu, J. Langhorst, C. Sina, R. Benamouzig, and J. Tack, "Whitepaper: Endomicroscopic Diagnosis of Food-Induced Allergy-Like Reactions," (2022), https://www.maunakeatech.com/landing/cellvizio-ibs-whitepaper.

5. M. Grover, A. Berumen, S. Peters, et al., "Intestinal Chemosensitivity in Irritable Bowel Syndrome Associates With Small Intestinal TRPV Channel Expression," *Alimentary Pharmacology & Therapeutics* 54, no. 9 (2021): 1179–1192, https://doi.org/10.1111/apt.16591.

6. R. Kiesslich, M. Goetz, E. M. Angus, et al., "Identification of Epithelial Gaps in Human Small and Large Intestine by Confocal Endomicroscopy," *Gastroenterology* 133, no. 6 (2007): 1769–1778, https://doi.org/10.1053/j.gastro.2007.09.011.

7. R. Kiesslich, H. Adib-Tezer, D. Teubner, et al., "Id: 3526039 Food Allergy Sensitivity Test (Fast) Withendomicroscopy of the Duodenum Enablestailored Exclusion Diet in Patients Withirritable Bowel Syndrome," *Gastrointestinal Endoscopy* 93 (2021): AB207.

8. R. Kiesslich, H. Adib-Tezer, D. Teubner, et al., "Su1344 Endomicroscopic Detection of Atypical Food Allergy in Patients With Irritable Bowel Syndrome – A New Diagnostic Era?," *Gastroenterology* 158, no. 6 (2020): S-558–S-559, https://doi.org/10.1016/S0016-5085(20)32099-0.

9. B. Gjini, I. Melchior, P. Euler, et al., "Food Intolerance in Patients With Functional Abdominal Pain: Evaluation Through Endoscopic Confocal Laser Endomicroscopy," *Endoscopy International Open* 11, no. 1 (2023): E67–E71, https://doi.org/10.1055/a-1978-6753.

10. T. Frieling, B. Gjini, I. Melchior, et al., "Gastrointestinal Adverse Reaction to Food (GARF) and Endoscopic Confocal Laser Endomicroscopy (eCLE)," *Zeitschrift für Gastroenterologie* 62, no. 08 (2024): 1201–1206, https://doi.org/10.1055/a-2258-8509.

11. Y. Q. Xiong, S. J. Ma, J. H. Zhou, X. S. Zhong, and Q. Chen, "A Meta-Analysis of Confocal Laser Endomicroscopy for the Detection of Neoplasia in Patients With Barrett's Esophagus," *Journal of Gastroenterology and Hepatology* 31, no. 6 (2016): 1102–1110, https://doi.org/10.1111/jgh.13267.

12. J. D. Machicado, B. Napoleon, A. M. Lennon, et al., "Accuracy and Agreement of a Large Panel of Endosonographers for Endomicroscopy-Guided Virtual Biopsy of Pancreatic Cystic Lesions," *Pancreatology* 22, no. 7 (2022): 994–1002, https://doi.org/10.1016/j.pan.2022.08.012.

13. J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement* 20, no. 1 (1960): 37–46, https://doi.org/10.1177/001316446002000104.

14. J. L. Fleiss, "Measuring Nominal Scale Agreement Among Many Raters," *Psychological Bulletin* 76, no. 5 (1971): 378–382, https://doi.org/10.1037/h0031619.

15. J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics* 33, no. 1 (1977): 159–174.

16. W. J. Youden, "Index for Rating Diagnostic Tests," *Cancer* 3, no. 1 (1950): 32–35, https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.

17. K. L. Gwet, *Handbook of Inter-Rater Reliability*, 4th ed. (Advanced Analytics, LLC, 2014).

18. M. Gamer, J. Lemon, I. Fellows, and S. Puspendra, "Package 'irr'," Published online January 26, 2019, accessed April 26, 2024, https://cran.r-project.org/web/packages/irr/irr.pdf.

19. X. Robin, N. Turck, A. Hainard, et al., "pROC: An Open-Source Package for R and S+ to Analyze and Compare ROC Curves," *BMC Bioinformatics* 12, no. 1 (2011): 77, https://doi.org/10.1186/1471-2105-12-77.

20. D. D. Childs, K. D. Hiatt, T. E. Craven, and J. J. Ou, "The Imaging Diagnosis of Cholecystitis in the Adult ED: A Comparative Multi-Reader, Multivariable Analysis of CT and US Image Features," *Abdominal Radiology* 47, no. 1 (2022): 184–195, https://doi.org/10.1007/s00261-021-03318-y.

Supporting Information

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