AMERICAN ACADEMY OF OPHTHALMOLOGY®

Reliability of Subjective Assessment of Spectral-Domain OCT Pathologic Features by Multiple Raters in Retinal Vein Occlusion

Sebastian Bemme, MD, Amelie Heins, DMD, Peer Lauermann, MD, Marcus Werner Storch, MD, Mohammed Haitham Khattab, MD, Hans Hoerauf, MD, Nicolas Feltgen, MD, Christian van Oterendorp, MD

Purpose: To examine the interrater and intrarater reliability of qualitatively and quantitatively assessed disorganization of retinal inner layers (DRIL) and disorganization of retinal outer layers (DROL) by multiple raters. Subjectively assessing these surrogate biomarkers can be challenging in daily routine, despite the high resolution of spectral-domain (SD) OCT scans.

Design: Retrospective trial.

Participants: Three hundred six pooled SD OCT scans of 34 patients treated for macular edema caused by retinal vein occlusion (RVO) between January 2016 and December 2017.

Methods: SD OCT scans were assessed by 6 raters regarding presence of cystoid macular edema, subretinal fluid (SRF), vitreoretinal traction, and epiretinal membrane and extent of DRIL and DROL.

Main Outcome Measures: Interrater and intrarater reliability were calculated applying κ statistics for qualitative assessment regarding each pathologic feature's presence in all evaluated OCT scans, and for quantified horizontal DRIL and DROL extent within each OCT cross-section.

Results: Cystoid macular edema and SRF assessments revealed excellent inter- and intrarater reliability with almost perfect strength of agreement, whereas subjective DRIL and DROL evaluations yielded low κ statistics with slight to moderate strength of agreement. Furthermore, the presence of SRF remarkably compromised the reliability of DROL detection.

Conclusions: Our data highlight the limited subjective assessibility of DRIL and DROL, underscoring the need for automated image analysis to improve the reliability of OCT biomarkers for clinical studies and daily practice. *Ophthalmology Science 2021;1:100031* © *2021 by the American Academy of Ophthalmology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).*

Supplemental material available at www.ophthalmologyscience.org.

Since OCT was introduced in ophthalmology, the in vivo visualization of individual retinal layers has improved greatly, to almost microscopic resolution. This development, together with its broad clinical application, spurred the evolution of numerous morphologic biomarkers, some of which indicate the absence of visual improvement despite the regression of subretinal or intraretinal fluid in the various retinal diseases associated with macular edema. Initially, the focus was mainly on the disorganization of retinal outer layers (DROL), or more precisely, on the disruption of the external limiting membrane (ELM), ellipsoid zone (EZ), and interdigitation zone (IZ).[1–10] Outer retinal tubulations, which represent invaginations of the photoreceptor layer, were identified as another pathologic feature of the retina's outer segment related to impaired visual function in age-related macular degeneration.[11–13] Recent investigations focused on the retina's inner segment: disorganization of retinal inner layers (DRIL) proved to be a negative predictor of visual outcome in diabetic macular edema (DME), retinal vein occlusion (RVO), central retinal artery occlusion, uveitis, and epiretinal membrane.[14–28]

The growing list of OCT pathologic characteristics encourages ophthalmologists to create a link between morphologic features and function, seeking to predict visual outcomes better before planned therapeutic interventions or to explain the lack of improvement in visual acuity after surgery. However, 2 main obstacles hinder the translation of established structure–function relationships into the daily practice of general ophthalmologists: first, the difficulty in detecting and interpreting numerous OCT pathologic features, and second, the inaccuracy in the definition of certain parameters, primarily DRIL. Even retina specialists may feel significant uncertainty regarding whether borderline areas with some anomaly in the layer's structure are DRIL.[29]

We hypothesize that among the various retinal pathologic features visible on OCT, this ambiguity is particularly pronounced for DRIL, but is rather low for cystoid macular edema (CME) and subretinal fluid (SRF). Consequently, the characterization of new OCT pathologic features would be accompanied by the need to develop and validate automated image analysis. To prove our hypothesis, we tested the reliability of subjective assessments of different

https://doi.org/10.1016/j.xops.2021.100031
ISSN 2666-9145/21

morphologic biomarkers visible on spectral domain (SD) OCT cross-sections in a cohort of patients with RVO.

Methods

Study Population and Image Acquisition

This study retrospectively enrolled 144 patients treated for newly diagnosed CME resulting from RVO between January 2016 and December 2017. It adhered to the tenets of the Declaration of Helsinki and was approved by the University Medical Center Göttingen institutional ethic committee (application no., 27/3/13). The requirement for informed consent was waived because of the retrospective nature of the study. Inclusion criteria required intravitreal anti−vascular endothelial growth factor therapy for treatment-naïve CME resulting from RVO, sufficiently reduced CME, and image acquisition via Spectralis SD OCT (Heidelberg Engineering) once before and at least once after anti−vascular endothelial growth factor treatment with 3 or fewer injections and no more than 6 months after CME was diagnosed. Exclusion criteria were insufficient image quality or pathologic features on OCT preventing reliable assessment of all retinal layers, such as significant hemorrhage-caused shadowing. Automated real-time tracking was applied in all scans. The number of averaged frames per OCT B-scan ranged from 7 to 21, with a median of 9 frames. Of the included OCT examinations, we analyzed only the central 3 horizontal OCT B-scans: 1 cutting the fovea, 1 above, and 1 below. Additionally, we assessed OCT B-scans of healthy fellow eyes. All OCT cross-sections were pooled, and 30 images were presented twice to calculate intrarater reliability.

Rating Characteristics

Six raters consisting of 3 consultants and 3 senior residents from our department, all experienced in assessing retinal SD OCT images, evaluated the OCT B-scans regarding the presence of the following pathologic features: CME, SRF, vitreoretinal traction, epiretinal membrane, DRIL, ELM disruption, EZ disruption, and IZ disruption.

All pathologic features were rated qualitatively, that is, whether a certain feature was either present or absent within the OCT B-scan, regardless of its extent. Disorganization of retinal inner layers and DROL were assessed quantitatively further, which means that in case of DRIL or disruption of ELM, EZ, or IZ (DROL) being rated as present, the rater had to mark the horizontal extent of the respective pathologic feature by a colored box within the OCT cross-section (Fig 1A). We requested a 2-level rating of DRIL: the red box had to extend over the retinal segment in which the rater was very certain about the presence of DRIL (DRIL certain), and the yellow box had to extend further over the segments where DRIL was suspected (DRIL suspected).

To reduce variability resulting from different individual concepts of DRIL and DROL, all raters were instructed thoroughly about these definitions before the project started in a joint training session. This included the presentation of reference images and discussion of the reference boxes drawn. Disorganization of retinal inner layers was defined as the inability to demarcate the ganglion cell−inner plexiform layer complex, the inner nuclear layer, and the outer plexiform layer against each other.[15,16] Disorganization of retinal outer layers was not assessed as a single feature, but rather was rated and analyzed separately for each outer retinal layer: ELM, EZ, and IZ. A to-be-marked DROL pathologic feature was defined as an obvious interruption of the respective layer not caused by a shadow artefact from overlying blood vessels.

MATLAB (MathWorks) was used to superimpose the marked pathologic features automatically in the respective OCT B-scan for all raters and to analyze the images further (Fig 1B).

Distribution of Ratings

First, we analyzed the qualitative assessment of all OCT scans by 6 raters (n = 6). For each OCT B-scan ($i = 1, 2, \ldots, N$; where $N = 306$), we counted the number of raters ($n_{i,\text{present}}$), who decided that a certain pathologic features were present within that OCT B-scan i. The distribution of how many scans had been rated as present for a certain pathologic feature by 0 to 6 raters then was calculated from that data. To assess interrater agreement separately regarding the presence and absence of a particular pathologic feature, we calculated p_{pres} and p_{abs}. Both parameters estimated the probability that most raters had assessed a sample exactly as 1 rater had rated it, that is, as present or absent for a given pathologic feature. In the qualitative assessment of OCT B-scans, p_{pres} represents the number of scans rated by a two-thirds majority of raters as present for a certain feature ($n_{i,\text{present}} \geq 4$) over the number of scans rated as present by at least 1 rater ($n_{i,\text{present}} \geq 1$):

$$p_{\text{pres}} = \frac{\textit{number of samples with } n_{i,\text{present}} \geq 4}{\textit{number of samples with } n_{i,\text{present}} \geq 1}. \quad (1)$$

P_{abs} represents the inverse approach, where the number of OCT B-scans in which a two-thirds majority did not detect a certain feature ($n_{i,\text{present}} \leq 2$) was set into relationship to the number of scans in which that feature had been rated absent by at least 1 rater ($n_{i,\text{present}} \leq 5$):

$$p_{\text{abs}} = \frac{\textit{number of samples with } n_{i,\text{present}} \leq 2}{\textit{number of samples with } n_{i,\text{present}} \leq 5}. \quad (2)$$

κ Statistics

Taking the standard approach assessing interrater and intrarater agreement, we further applied κ statistics from the observed percentage of agreement (p_0) and probability of agreement by chance (p_e), regarding the 2 categories (absent or present) of each OCT pathologic feature.

$$\kappa = \frac{p_0 - p_e}{1 - p_e}. \quad (3)$$

To calculate interrater reliability, we applied Fleiss' κ (κ_F) value because OCT B-scans were assessed by multiple raters.[30] The κ_F value was interpreted as strength of agreement from poor to almost perfect according to Landis and Koch.[31] To test for intrarater reliability, 30 OCT B-scans from the image pool were rated and marked twice in blinded fashion. Intrarater reliability then was assessed by calculating Cohen's κ (κ_C) value between repeatedly rated scans separately for each rater.[32,33] Supplemental Table 1 provides details on our calculation of p_0 and p_e for κ_F and κ_C.

Quantitative Assessment

Besides the reliability of the qualitative assessment, which disregarded the extent of the marked pathologic feature, we additionally evaluated the interrater and intrarater reliability of the DRIL and DROL quantitative assessments. From the images including all raters' marks as in Figure 1B, we analyzed each vertical pixel column ($i = 1, \ldots, N$; where N = entire horizontal B-scan length: 1024 in 296 OCT B-scans, 5013 in 7 OCT B-scans, and 1536 in 3 OCT B-scans) as a separate sample. We counted the number of raters ($n_{i,\text{present}}$), who marked the respective vertical pixel column as present for that pathologic feature. This was carried out automatically and separately for each OCT B-scan with all raters' superimposed markings using

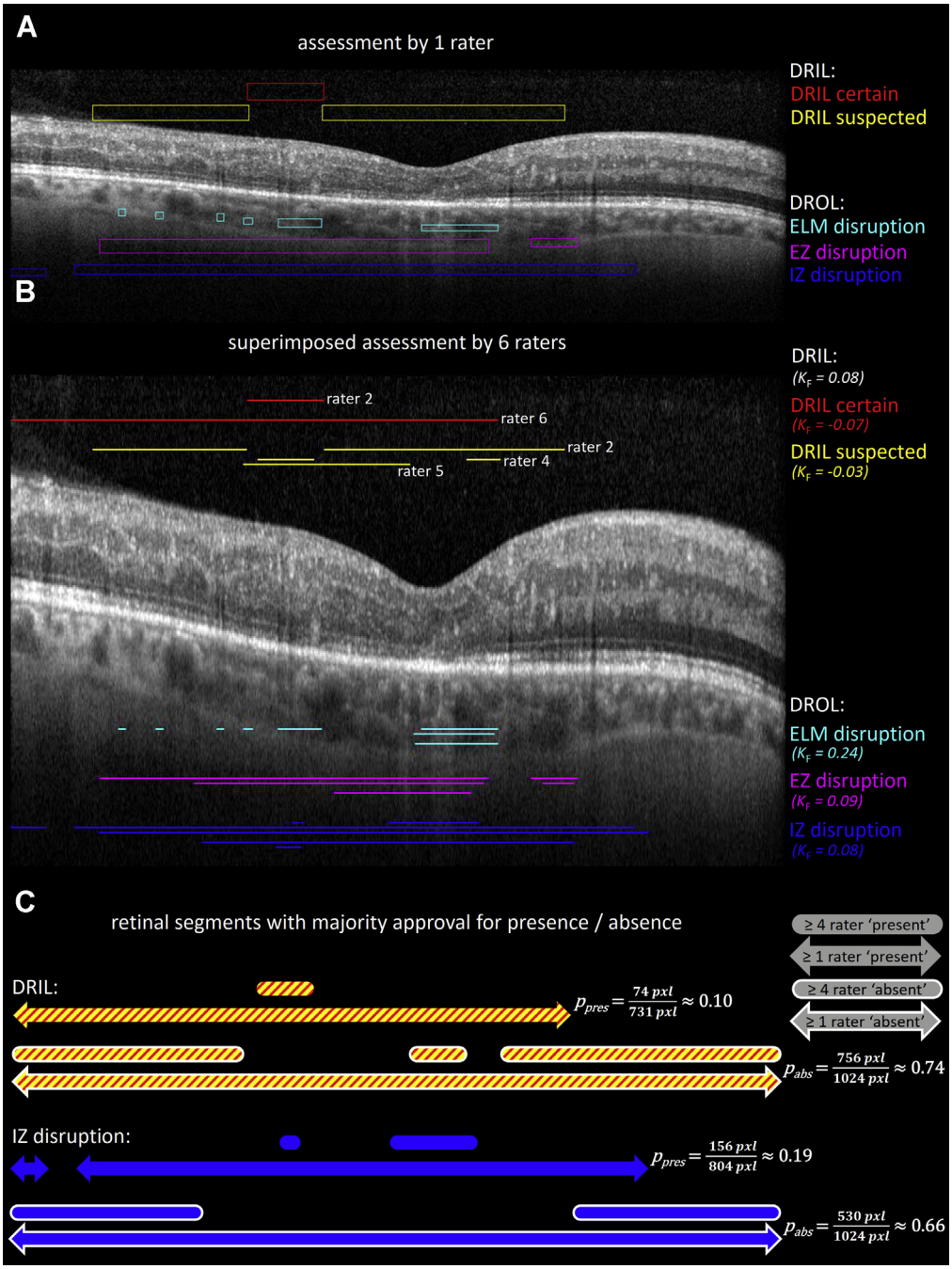

Figure 1. Spectral-domain OCT assessed for disorganization of retinal inner layers (DRIL) and disorganization of retinal outer layers (DROL). **A**, OCT B-scan as assessed by 1 rater with colored markings of different pathologic features. Disorganization of retinal inner layers was marked with a red or yellow box, depending on whether the rater was very certain about the presence of DRIL (DRIL certain) or DRIL was just suspected (DRIL suspected), respectively. Disorganization of retinal outer layers was assessed separately for each outer retinal layer: external limiting membrane (ELM; cyan), ellipsoid zone (EZ; magenta), and interdigitation zone (IZ; blue). The vertical extent of the colored boxes was irrelevant. **B**, Same OCT B-scan with all raters' colored markings superimposed (vertically stretched for better visualization of the superimposed lines). **C**, Schematic illustration of the linear extent of retinal segments of the OCT B-scan in (**A**) and (**B**) with majority approval for presence and absence to illustrate retinal segments with good agreement versus those with low agreement for DRIL and IZ disruption in the example. Relatively short segments rated as present by 4 raters or more (rounded ends, no frame) over the length of the segment rated as present by at least 1 rater (arrowhead ends, no frame) resulted in low probability of majority approval for presence (p_{pres}). In contrast, horizontal extent of the segments rated as absent (white frame) by 1 rater or more and 4 raters or more were less different, resulting in higher probability of majority approval for absence (p_{abs}). pxl = pixels.

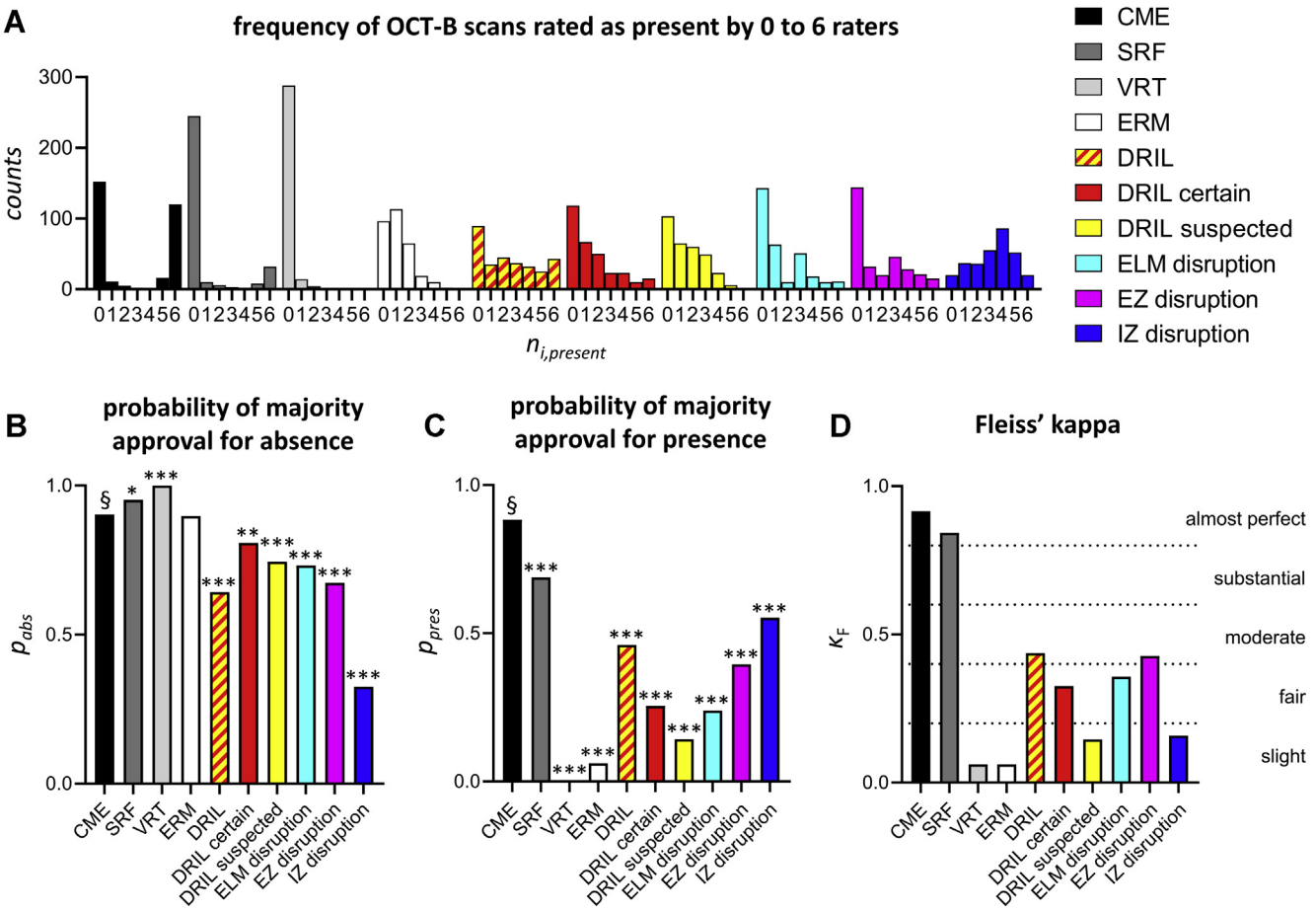

Figure 2. Bar graphs showing interrater reliability of qualitatively assessed OCT B-scans. **A**, frequency distribution of all OCT B-scans rated present for the respective pathologic feature by 0 to 6 raters ($n_{i,present} = 0,1, \ldots ,6$). $n_{i,present} = 0$ means that all raters assessed the respective pathologic feature as absent and $n_{i,present} = 6$ means that all raters assessed the respective pathologic feature as present. **B**, **C**, Probability of majority approval (≥ 4 raters) for a single rater's decision on (**A**) the absence (p_{abs}) or (**B**) the presence (p_{pres}) of a certain pathologic feature. Asterisks indicate a significant difference (*$P < 0.05$, **$P < 0.01$, and ***$P < 0.001$) to p_{abs} and to p_{pres} of CME (§) applying the chi-square test. **D**, Fleiss' κ (κ_F) and strength of agreement. CME = cystoid macular edema; DRIL = disorganization of retinal inner layers; ELM = external limiting membrane; ERM = epiretinal membrane; EZ = ellipsoid zone; IZ = interdigitation zone; SRF = subretinal fluid; VRT = vitreoretinal traction.

MATLAB software. The p_{pres} and p_{abs} values as well as the κ_F value then were calculated for each OCT B-scan separately. Here, p_{pres} and p_{abs} represent the cumulative segment length where most raters marked a certain feature as present ($n_{i,present} \geq 4$) or absent ($n_{i,present} \leq 2$) relative to the scan length that had been marked as present ($n_{i,present} \geq 1$) or absent ($n_{i,present} \leq 5$), respectively, by at least 1 rater (illustrated for DRIL and IZ disruption in Fig 1C). OCT cross sections marked twice evaluated the intrarater reliability of quantitative assessments of DRIL and DROL calculating the κ_C value for each rater.

Results

Population Characteristics and Image Pool

Thirty-four patients, 12 women and 22 men with a mean age of 67 years (range, 35–83 years) were included. Three fovea-centered horizontal OCT B-scans from 76 OCT examinations of eyes with RVO and from 16 OCT examinations of healthy fellow eyes, together with 30 repeatedly presented OCT B-scans, formed our image pool (n = 306) analyzed by 6 raters.

Interrater Reliability

Qualitative Assessment. Figure 2A displays the distribution of OCT B-scans qualitatively rated as present for the respective pathologic feature by 0 to 6 raters. Perfect interrater reliability results in a so-called bipolar pattern of distribution, with all counts being either $n_{i,present} = 0$ or $n_{i,present} = 6$ and 0 counts for $n_{i,present} = 1$ to 5. For all pathologic features except epiretinal membrane and IZ disruption, the frequency of OCT B-scans with $n_{i,present} = 0$ considerably exceeded the frequencies of OCT B-scans rated as present by at least 1 rater. However, OCT B-scans with $n_{i,present} = 6$ yielded the second most frequent proportion only concerning CME and SRF.

We observed a majority approval for absence of CME, SRF, vitreoretinal traction, and epiretinal membrane in 90% to 100% of those OCT B-scans that had been rated as absent

Table 1. Interrater Reliability

Spectral-Domain OCT Pathologic Feature	Cystoid Macular Edema	Subretinal Fluid	Vitreoretinal Traction	Epiretinal Membrane	Disorganization of Retinal Inner Layers	Disorganization of Retinal Inner Layers Rating		Disruption		
						Certain	*Suspected*	*External Limiting Membrane*	*Ellipsoid Zone*	*Interdigitation Zone*
Qualitative assessment										
p_{abs}	0.90	0.95	1.00	0.90	0.64	0.81	0.75	0.73	0.67	0.33
p_{pres}	0.88	0.69	0.00	0.06	0.46	0.26	0.14	0.24	0.40	0.55
κ_F	0.92	0.84	0.06	0.06	0.44	0.33	0.15	0.36	0.43	0.16
Strength of agreement	Almost perfect	Almost perfect	Poor to slight	Poor to slight	Moderate	Fair	Slight	Fair	Moderate	Slight
Quantitative assessment										
p_{abs} (mean ± SD)					0.89 ± 0.20	0.94 ± 0.14	0.99 ± 0.04	0.94 ± 0.14	0.94 ± 0.13	0.79 ± 0.21
p_{pres} (mean ± SD)					0.16 ± 0.26	0.08 ± 0.18	0.00 ± 0.03	0.08 ± 0.18	0.13 ± 0.22	0.14 ± 0.20
κ_F (mean ± SD)					0.11 ± 0.21	0.02 ± 0.17	0.01 ± 0.08	0.12 ± 0.19	0.21 ± 0.23	0.09 ± 0.20
Strength of agreement					Slight	Slight	Slight	Slight	Fair	Slight

p_{abs} = probability of majority approval for absence; p_{pres} = probability of majority approval for presence; κ_F = Fleiss' κ; SD = standard deviation.

for that pathologic feature at least by 1 rater (p_{abs}; Table 1; Fig 2B). The percentage of majority approval on OCT scans that were rated as present by at least 1 rater was highest for CME with 88%, followed by SRF with 69% (p_{pres}; Table 1; Fig 2C). Consequently, the interrater reliability for CME and SRF as calculated by κ statistics revealed almost perfect strength of agreement as indicated by $\kappa_F = 0.92$ and $\kappa_F = 0.84$, respectively (Table 1; Fig 2D).

In contrast, the interrater reliability regarding retinal layer disruption (DRIL, ELM, EZ, or IZ) was markedly lower, with the κ_F value ranging between 0.16 and 0.44, indicating only slight to moderate strength of agreement (Table 1; Fig 2D). Although p_{abs} for the DRIL assessment and disruption of ELM and EZ was relatively high (p_{abs}, 0.64−0.81), p_{pres} regarding those pathologic features yielded low values of between 0.14 and 0.46. The probability of majority approval was higher for presence ($p_{pres} = 0.55$) than for absence ($p_{abs} = 0.33$) regarding IZ disruption.

Quantitative Assessment of Retinal Layer Disruption. Calculating the interrater reliability of the pathologic features' exact localization and extent within each OCT B-scan (DRIL, ELM, EZ, and IZ) yielded evidence similar to the qualitative assessment (Fig 3A−C). The κ statistics showed slight to fair strength of agreement on average (mean of κ_F, 0.01−0.21; Fig 3C). Although most of the scans exhibited high probability of majority approval for absence (p_{abs}, 0.79−0.99; Fig 3A), the probability of majority approval for presence was very low (p_{pres}, 0.00−0.16 for all layer disruptions; Fig 3B). We validated the concept of majority approval by correlating the product of $p_{pres} \times p_{abs}$ with κ_F (Supplemental Fig 1), showing good consistency for both approaches ($R^2 > 0.73$; $P < 0.0001$) except for quantitatively assessed DRIL suspected ($R^2 = 0.21$).

Furthermore, we conducted a pairwise correlation analysis of the length of the marked pathologic features' horizontal extent for all possible rater pairs (Fig 3D). In contrast to the aforementioned κ statistics, this analysis did not consider the marks' exact localizations, only their total length. Pearson's correlation coefficients (r) ranged from 0.03 to 0.95, with a median of 0.70, 0.18, 0.23, and 0.15 for DRIL, ELM disruption, EZ disruption, and IZ disruption, respectively.

Intrarater Reliability

Repeated qualitative assessments by the raters yielded excellent intrarater reliability only when assessing CME, with κ_C values of between 0.85 and 1.00. The strength of agreement was less on average for all other pathologic features, demonstrating relatively broad variability with κ_C values ranging from −0.03 to 1.00 (Fig 4A). The mean κ_C value of all raters regarding DRIL certain and IZ disruption was significantly lower than for CME (mean ± standard deviation: CME, 0.93 ± 0.07; DRIL certain, 0.56 ± 0.26; IZ disruption, 0.57 ± 0.23; $P = 0.027$ [CME vs. DRIL certain] and $P = 0.038$ [CME vs. IZ disruption], paired t test with Bonferroni-Holm correction). The strength of agreement of the quantitative assessment, that is, of a repeatedly marked horizontally extended layer

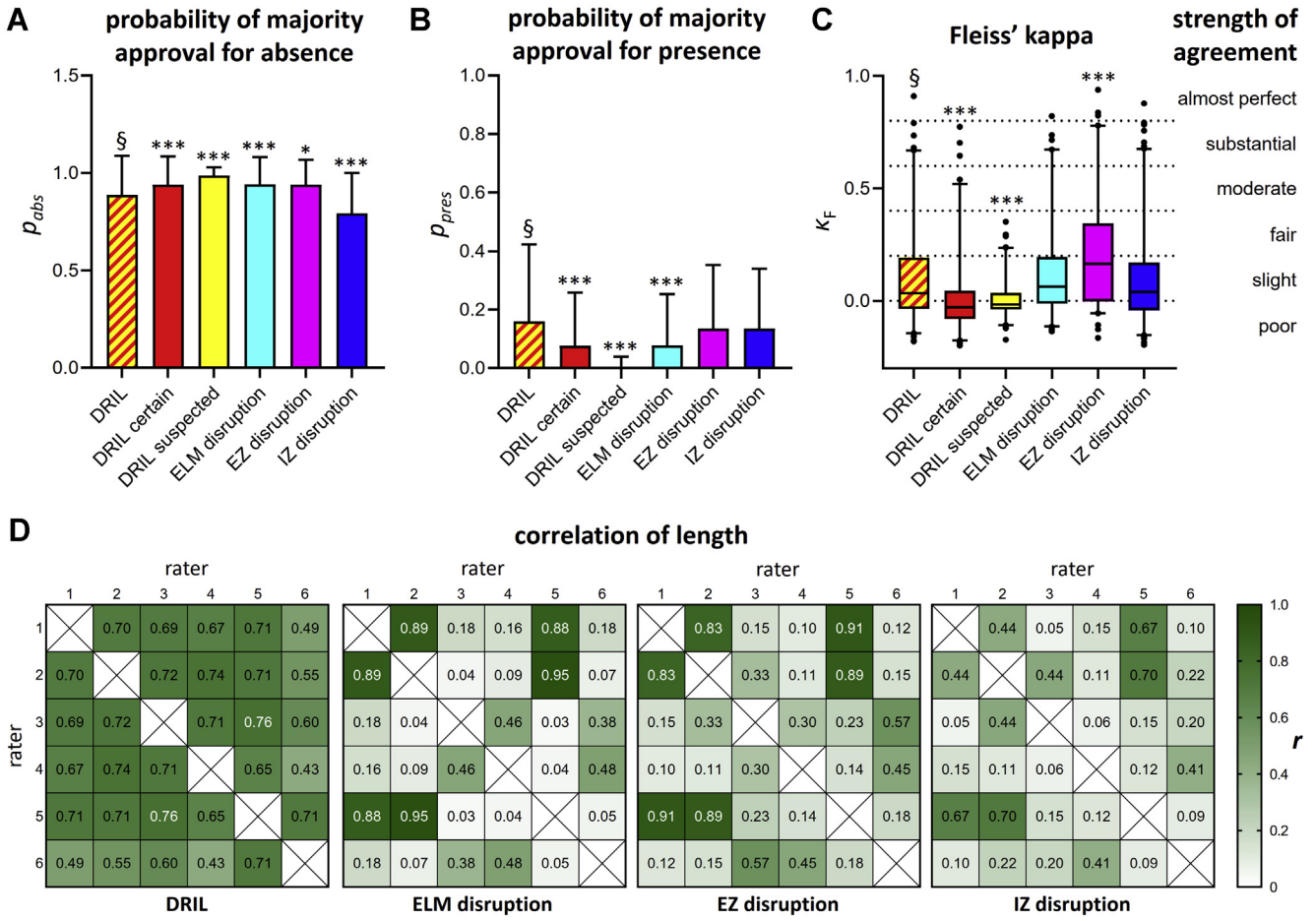

Figure 3. Graphs and matrices showing interrater reliability of quantitative assessments of layer disruptions. **A**, **B**, Bar graphs showing the probability of majority approval (≥ 4 raters) for a single rater's decision on (**A**) the absence (p_{abs}) or (**B**) the presence (p_{pres}) of a certain layer disruption. The calculation was performed for each vertical pixel column of each OCT B-scan. Thus, the columns and error bars represent mean ± standard deviation values for all OCT scans. **C**, Box-and-whisker plot showing Fleiss' κ (κ_F) value and corresponding strength of agreement for the quantitative assessment. The box-and-whisker plot displays the median and range from the first to third quartile by a line and a box, respectively, with whiskers indicating the 2.5% and 97.5% percentiles. **D**, Correlation matrices of pairwise correlation of the marked pathologic feature's length. Numbers and color coding show Pearson's correlation coefficient. Asterisks (**A**–**C**) indicate a significant difference (*$P < 0.05$, **$P < 0.01$, and ***$P < 0.001$) to p_{abs}, p_{pres}, and κ_F of DRIL (§) applying the Mann–Whitney *U* test (p_{abs} and p_{pres}) and *t* test (κ_F), respectively. DRIL = disorganization of retinal inner layers; ELM = external limiting membrane; EZ = ellipsoid zone; IZ = interdigitation zone.

disruption, also exhibited considerable variability. Notably, the DRIL was marked less consistently at repeated assessments than outer retinal layer disruptions (ELM, EZ, IZ), with the difference between DRIL certain and IZ disruption being statistically significant (mean ± standard deviation κ_C value: DRIL certain, 0.31 ± 0.15 vs. IZ disruption, 0.49 ± 0.2; $P = 0.036$, paired *t* test).

Influencing Factors on Interrater Reliability

Strength of agreement regarding DRIL and DROL assessed by different raters ranged widely from poor to substantial within the 2.5% to 97.5% percentile (Fig 3C). Therefore, we aimed to identify factors influencing the interrater reliability of DRIL and DROL, hypothesizing that the presence of CME and SRF might have played a significant role. Thus, we compared κ_F in OCT B-scans with CME ($n_{i,present} = 6$), with SRF ($n_{i,present} = 6$), and without CME and SRF ($n_{i,present} = 0$). The κ_F value of qualitatively assessed ELM, EZ, and IZ disruption was markedly lower when CME was present and lowest when SRF was present, whereas the κ_F value of DRIL assessment was almost equal among the 3 groups (Fig 5A1). As for quantitative assessment, agreement of horizontal EZ disruption extent was significantly lower with CME and SRF than without (Fig 5A2). Interestingly, agreement of horizontal DRIL extent was significantly stronger in scans with CME or SRF (Fig 5A2). Regarding the impact of image quality, we found that automated real-time tracking and signal quality failed to correlate significantly with the κ_F value of quantitatively assessed OCT B-scans.

Another conceivable influencing factor is the raters' amount of clinical experience. Therefore, we compared the

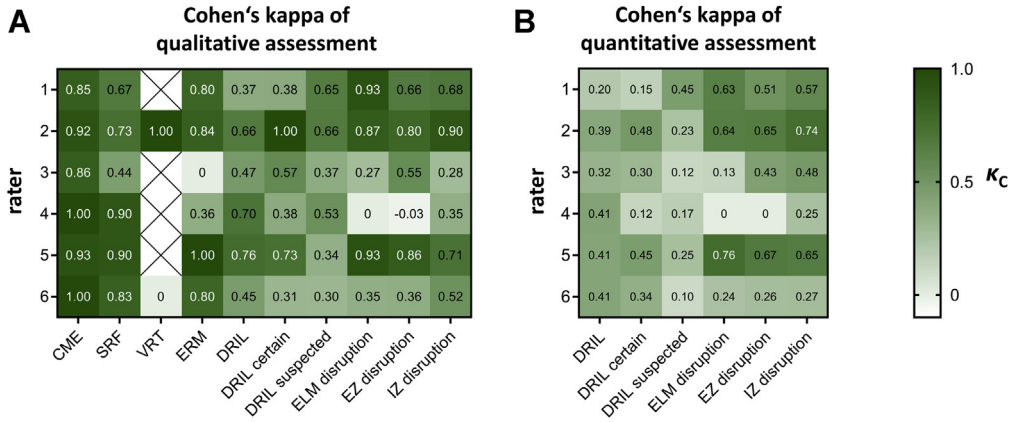

Figure 4. Correlation matrices showing intrarater reliability calculated as Cohen's κ (κ_C) value. The κ_C value was calculated from 30 OCT B-scans, which were rated and marked twice. The κ_C values are displayed for each rater and each pathologic feature as color-coded tiles ranging from white to green (−0.1 to 1.0). The κ_C value was calculated for (**A**) the qualitative assessment as well as for (**B**) quantitative assessments of the extent of layer disruptions. For qualitative assessments, the mean κ_C value of all raters was significantly higher for cystoid macular edema (CME) compared with disorganization of retinal inner layers (DRIL) certain and interdigitation zone (IZ; $P = 0.027$ and $P = 0.038$, respectively, paired t test with Bonferroni-Holm correction). As for the quantitative assessment, the consistency of the extent of DRIL certain was significantly lower than for IZ disruption ($P = 0.036$, paired t test). ELM = external limiting membrane; ERM = epiretinal membrane; EZ = ellipsoid zone; SRF = subretinal fluid; VRT = vitreoretinal traction.

agreement between consultants (n = 3) and residents (n = 3), which yielded slightly but still significantly better interrater reliability among consultants (Fig 5B1, B2).

Discussion

The improved visualization of individual retinal layers through continuously advancing OCT technology has been accompanied by the evolution of various morphologic biomarkers going far beyond the detection of intraretinal or subretinal fluid. However, more advanced OCT biomarkers such as DRIL and DROL may be harder to detect, and quantifying their extent may be impeded by the considerable ambiguity of anomalies observed on OCT. We therefore hypothesized that the interrater and intrarater reliability of subjective ratings of DRIL and DROL would be lower than those of CME and SRF.

Indeed, our data yielded excellent interrater and intrarater reliability for well-known pathologic features like CME and SRF. The same applied for a healthy retina, indicated by the high probability of majority approval for the absence of almost all pathologic features (except IZ disruption). However, the assessments of layer disruptions, including both DRIL and DROL, revealed only moderate strength of interrater and intrarater agreement. A few studies have reported on DRIL's interrater reliability, but their comparability is limited by various factors such as a purely qualitative assessment (DRIL absent or present), the size of the retinal segment chosen for assessment, and the definition of DRIL. One recently published trial concurring with our findings reported only slight to moderate agreement of qualitative DRIL assessments.[17,29] In contrast, other trials reported good agreement.[14–16,19,23,25] However, targeting the association between DRIL and visual acuity, most evaluated only a foveally centered zone with a diameter of 1000 or 1500 μm,[14–16,23,25] whereas in our study, similar to Babiuch et al,[17] we assessed the entire OCT B-scan. As for the DRIL definition, we adopted the established concept of the inability to identify or demarcate the boundaries between the ganglion cell–inner plexiform layer complex, inner nuclear layer, and outer plexiform layer.[15,16] In addition, some studies set certain thresholds, like more than 50% foveal-center involvement, or a more than 20 μm DRIL extent.[15,16] For the qualitative assessment of DRIL in our study, we set no such thresholds, but the smallest extent of marked DRIL in our study was not less than 100 μm.

We also analyzed the interrater and intrarater agreement of quantitative assessments of DRIL and DROL. By superimposing all raters' marks, our analysis applying κ statistics considered the agreement regarding the extent and localization of the respective pathologic feature. Here, the strength of agreement between raters regarding DRIL was even worse, ranging mostly (25%–75% quartile) at the slight agreement level. Other trials that quantitatively measured DRIL in RVO and DME reported good agreement. However, they correlated DRIL lengths when assessing agreement (thus failing to consider the spatial overlay of assessed DRIL); this factor possibly caused the better agreement.[14,25] In fact, mimicking this approach by a pairwise correlation of DRIL length in our study yielded a median Pearson's correlation coefficient of 0.7, suggesting good agreement. Other studies measured the DRIL extent repeatedly until the intergrader correlation was satisfactory or until reaching a consensus on disagreements.[26,28]

Assessments of outer retinal layer disruption (ELM, EZ, IZ) also yielded only slight to moderate strength of agreement in our study, with the lowest κ_F value for judging IZ disruption. Those findings of ours contradict those of other studies reporting good intergrader reliability assessing EZ and ELM disruption. However, they used a different methodology calculating agreement only within the foveally

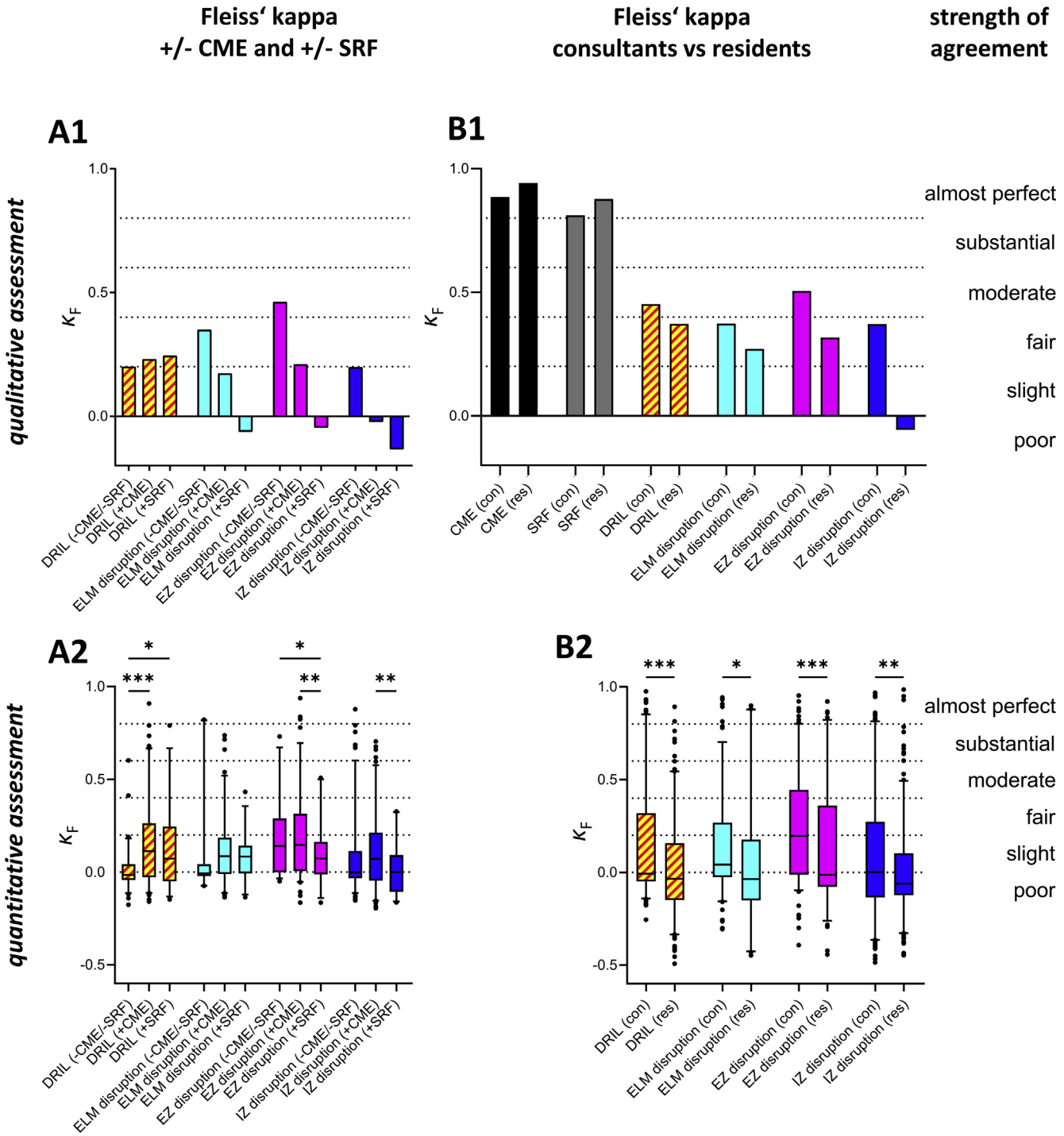

Figure 5. Bar graphs and box-and-whisker plots showing interrater reliability of qualitatively and quantitatively assessed OCT B-scans in dependence of coexisting pathologic features and clinical experience. **A1**, **A2**, Fleiss' κ (κ_F) value regarding assessment of disorganization of retinal inner layers (DRIL) and disorganization of retinal outer layers (DROL) in the group of OCT B-scans without cystoid macular edema (CME) and subretinal fluid (SRF) was compared with κ_F of DRIL and DROL assessment in the 2 groups of scans with CME and with SRF. **B1**, **B2**, κ_F values of OCT evaluation by consultants compared with the κ_F value of residents' assessment. **A2**, **B2**, Box-and-whisker plots displaying the median and range from the first to the third quartile by a line and a box, respectively, with whiskers indicating the 2.5% and 97.5% percentile. Asterisks indicate significant difference between compared groups (*$P < 0.05$, **$P < 0.01$, and ***$P < 0.001$; t test). ELM = external limiting membrane; EZ = ellipsoid zone; IZ = interdigitation zone.

centered 1000-μm zone.[1,2] Regarding the number of raters, the general standard has been evaluation by 2 masked retinal specialists, plus a third one in case of disagreement. However, some studies did not calculate or state interrater reliability, nor were OCT scans assessed by multiple observers.[34,35]

Intraclass correlation is an alternative method for analyzing the agreement of interval-scaled parameters measured by multiple observers. Therefore, intraclass correlation was applied occasionally when more than 2 graders assessed the length of disrupted retinal boundaries on OCT cross sections.[19] However, our data, especially the widely ranging Person's r values of the pairwise correlated DROL extent, failed to indicate any exchangeability, a prerequisite for intraclass correlation. Interestingly, some raters seemed to harmonize in unison when assessing the length of outer retinal layers, especially ELM and EZ disruption (Fig 3D), whereas other raters disagreed completely. This may indicate a similar approach to interpreting OCT scans, however, only within a certain subgroup of our raters. In our single-center design, we cannot completely exclude that this subgroup was influenced by social factors such as close-colleague or mentor–mentee combinations in the clinical routine. Despite this certain interdependence between some raters, overall agreement regarding DRIL and DROL was still only slight to moderate. A multicenter study would ensure a higher level of independence between raters.

The presence of various copathologies may impede the assessment of retinal layer disruptions.[29] We acknowledge that intraretinal and subretinal fluid markedly compromised interrater agreement regarding the qualitative assessment of DROL, but not DRIL. In particular, the strength of agreement regarding ELM, EZ, and IZ disruption was worst in OCT cross sections with subretinal fluid. For the quantitative assessment, the κ_F value of EZ disruption was significantly lower in OCT B-scans with SRF than in those without CME and SRF. Our data showed that assessments of photoreceptor integrity before macular edema has resolved, which should be interpreted with caution. Consequently, Shin et al[1] assessed EZ and ELM integrity only at the final visit after DME resolution. Regarding DRIL: our study's raters demonstrated significantly higher agreement over DRIL when CME or SRF was present. We hypothesize that this was because we had assigned DRIL quite consistently to areas where CME was present in the inner retina, perhaps caused by increased false-positive DRIL ratings biased by the copathology. Thus, the question remains regarding how reliable DRIL detection can ever be in the presence of pathologic features like CME and SRF. Most studies assessed DRIL despite the presence of CME. However, do invisible layer demarcations in the presence of cystoid spaces actually represent immediate deterioration of the retinal network's layered architecture also on the microscopic scale, or does CME merely impede the identification of tissue borders of different reflectivity? Radwan et al[14] characterized DRIL resolution patterns thoroughly in patients with DME and showed no significant difference in visual acuity improvement in those with late and early DRIL resolution compared with no baseline DRIL, evidence that may support the second hypothesis mentioned above.

The limited reliability of the subjective assessment of retinal layer disruption demonstrated in our study has a significantly negative impact on clinical studies testing the relevance of these biomarkers. Moreover, the considerable ambiguity and room for personal interpretation—which pertains to DRIL in particular—hinders its usefulness and transfer to the daily practice of ophthalmologists. In our opinion, this highlights the need for establishing objective methods to detect layer disruption. Naturally, such methods would not necessarily yield across-the-board accurate judgements on the presence or absence of retinal layer disruptions. However, they could enable the application of a shared standard for ophthalmologists, which in turn would mean greater consistency in DRIL detection across clinical studies and in clinical application. One potential approach is to develop and validate automated or semiautomated image analysis. For example, Sun et al[15] measured EZ and ELM reflectivity in addition to subjective assessments, and Itoh et al[36] introduced volumetric EZ mapping. Machine learning-based algorithms already have proven to be valuable approaches for the automated detection of anomalies in the outer retina.[37,38]

Conclusions

Compared with the excellent interrater and intrarater reliability of subjectively assessed CME and SRF, DRIL and DROL evaluated by multiple raters yielded only slight to moderate strength of agreement. The limited subjective assessibility of inner and outer retinal layer disorganization underscores the need for automated image analysis, which would facilitate both reliable OCT classifications for clinical studies and the adoption of advanced OCT biomarkers in daily practice.

Footnotes and Disclosures

Originally received: February 18, 2021.
Final revision: May 28, 2021.
Accepted: June 1, 2021.
Available online: June 5, 2021. Manuscript no. D-21-00014.

Department of Ophthalmology, University Medical Center Göttingen, Göttingen, Germany.

Disclosure(s):
All authors have completed and submitted the ICMJE disclosures form.
The author(s) have no proprietary or commercial interest in any materials discussed in this article.
Financially supported by Novartis Pharma GmbH, Nuremberg, Germany.

HUMAN SUBJECTS: Human subjects were included in this study. The human ethics committees at the University Medical Center Göttingen approved the study. All research adhered to the tenets of the Declaration of Helsinki. The requirement for informed consent was waived because of the retrospective nature of the study.

No animal subjects were included in this study.

Author Contributions:

Conception and design: Bemme, Hoerauf, Feltgen, van Oterendorp

Analysis and interpretation: Bemme, van Oterendorp

Data collection: Bemme, Heins, Lauermann, Storch, Khattab, Feltgen, van Oterendorp

Obtained funding: N/A; Study was performed as part of regular employment duties at Department of Ophthalmology, University Medical Center Göttingen. No additional funding was provided.

Overall responsibility: Bemme, Hoerauf, Feltgen, van Oterendorp

Abbreviations and Acronyms:
CME = cystoid macular edema; **DME** = diabetic macular edema; **DRIL** = disorganization of retinal inner layers; **DROL** = disorganization of retinal outer layers; **ELM** = external limiting membrane; **ERM** = epiretinal membrane; **EZ** = ellipsoid zone; **IZ** = interdigitation zone; κ_C = Cohen's κ; κ_F = Fleiss' κ; **RVO** = retinal vein occlusion; **SD** = spectral-domain; **SRF** = subretinal fluid; **VRT** = vitreoretinal traction.

Keywords:
RVO, SD-OCT, interrater reliability, DRIL, DROL.

Correspondence:
Sebastian Bemme, MD, Department of Ophthalmology, University Medical Center Göttingen, Robert-Koch-Str, 4037075 Göttingen, Germany. E-mail: Sebastian.bemme@med.uni-goettingen.de.

References

1. Shin HJ, Lee SH, Chung H, Kim HC. Association between photoreceptor integrity and visual outcome in diabetic macular edema. *Graefes Arch Clin Exp Ophthalmol*. 2012;250(1):61−70.
2. Shin HJ, Chung H, Kim HC. Association between integrity of foveal photoreceptor layer and visual outcome in retinal vein occlusion. *Acta Ophthalmol*. 2011;89(1):e35−e40.
3. Piccolino FC, de la Longrais RR, Ravera G, et al. The foveal photoreceptor layer and visual acuity loss in central serous chorioretinopathy. *Am J Ophthalmol*. 2005;139(1):87−99.
4. Ota M, Tsujikawa A, Murakami T, et al. Foveal photoreceptor layer in eyes with persistent cystoid macular edema associated with branch retinal vein occlusion. *Am J Ophthalmol*. 2008;145(2):273−280.
5. Ota M, Tsujikawa A, Murakami T, et al. Association between integrity of foveal photoreceptor layer and visual acuity in branch retinal vein occlusion. *Br J Ophthalmol*. 2007;91(12):1644−1649.
6. Ota M, Tsujikawa A, Kita M, et al. Integrity of foveal photoreceptor layer in central retinal vein occlusion. *Retina*. 2008;28(10):1502−1508.
7. Murakami T, Tsujikawa A, Ohta M, et al. Photoreceptor status after resolved macular edema in branch retinal vein occlusion treated with tissue plasminogen activator. *Am J Ophthalmol*. 2007;143(1):171−173.
8. Maheshwary AS, Oster SF, Yuson RM, et al. The association between percent disruption of the photoreceptor inner segment-outer segment junction and visual acuity in diabetic macular edema. *Am J Ophthalmol*. 2010;150(1):63−67.e1.
9. Eandi CM, Chung JE, Cardillo-Piccolino F, Spaide RF. Optical coherence tomography in unilateral resolved central serous chorioretinopathy. *Retina*. 2005;25(4):417−421.
10. Kang HM, Chung EJ, Kim YM, Koh HJ. Spectral-domain optical coherence tomography (SD-OCT) patterns and response to intravitreal bevacizumab therapy in macular edema associated with branch retinal vein occlusion. *Graefes Arch Clin Exp Ophthalmol*. 2013;251(2):501−508.
11. Lee JY, Folgar FA, Maguire MG, et al. Outer retinal tubulation in the Comparison of Age-Related Macular Degeneration Treatments Trials (CATT). *Ophthalmology*. 2014;121(12):2423−2431.
12. Faria-Correia F, Barros-Pereira R, Queiros-Mendanha L, et al. Characterization of neovascular age-related macular degeneration patients with outer retinal tubulations. *Ophthalmologica*. 2013;229(3):147−151.
13. Zweifel SA, Engelbert M, Laud K, et al. Outer retinal tubulation: a novel optical coherence tomography finding. *Arch Ophthalmol*. 2009;127(12):1596−1602.
14. Radwan SH, Soliman AZ, Tokarev J, et al. Association of disorganization of retinal inner layers with vision after resolution of center-involved diabetic macular edema. *JAMA Ophthalmol*. 2015;133(7):820−825.
15. Sun JK, Radwan SH, Soliman AZ, et al. Neural retinal disorganization as a robust marker of visual acuity in current and resolved diabetic macular edema. *Diabetes*. 2015;64(7):2560−2570.
16. Mimouni M, Segev O, Dori D, et al. Disorganization of the retinal inner layers as a predictor of visual acuity in eyes with macular edema secondary to vein occlusion. *Am J Ophthalmol*. 2017;182:160−167.
17. Babiuch AS, Han M, Conti FF, et al. Association of disorganization of retinal inner layers with visual acuity response to anti-vascular endothelial growth factor therapy for macular edema secondary to retinal vein occlusion. *JAMA Ophthalmol*. 2019;137(1):38−46.
18. Sun JK, Lin MM, Lammer J, et al. Disorganization of the retinal inner layers as a predictor of visual acuity in eyes with center-involved diabetic macular edema. *JAMA Ophthalmol*. 2014;132(11):1309−1316.
19. Zur D, Iglicki M, Feldinger L, et al. Disorganization of retinal inner layers as a biomarker for idiopathic epiretinal membrane after macular surgery-the DREAM Study. *Am J Ophthalmol*. 2018;196:129−135.
20. Yilmaz H, Durukan AH. Disorganization of the retinal inner layers as a prognostic factor in eyes with central retinal artery occlusion. *Int J Ophthalmol*. 2019;12(6):990−995.
21. Ishibashi T, Sakimoto S, Shiraki N, et al. Association between disorganization of retinal inner layers and visual acuity after proliferative diabetic retinopathy surgery. *Sci Rep*. 2019;9(1):12230.
22. Grewal DS, O'Sullivan ML, Kron M, Jaffe GJ. Association of disorganization of retinal inner layers with visual acuity in eyes with uveitic cystoid macular edema. *Am J Ophthalmol*. 2017;177:116−125.
23. Garnavou-Xirou C, Xirou T, Gkizis I, et al. The role of disorganization of retinal inner layers as predictive factor of postoperative outcome in patients with epiretinal membrane. *Ophthalmic Res*. 2020;63(1):13−17.
24. Busch C, Okada M, Zur D, et al. Baseline predictors for visual acuity loss during observation in diabetic macular oedema with good baseline visual acuity. *Acta Ophthalmol*. 2020;98(7):e801−e806.
25. Berry D, Thomas AS, Fekrat S, Grewal DS. Association of disorganization of retinal inner layers with ischemic index and visual acuity in central retinal vein occlusion. *Ophthalmol Retina*. 2018;2(11):1125−1132.

26. Joltikov KA, Sesi CA, de Castro VM, et al. Disorganization of retinal inner layers (DRIL) and neuroretinal dysfunction in early diabetic retinopathy. *Invest Ophthalmol Vis Sci*. 2018;59(13):5481−5486.
27. Das R, Spence G, Hogg RE, et al. Disorganization of inner retina and outer retinal morphology in diabetic macular edema. *JAMA Ophthalmol*. 2018;136(2):202−208.
28. Kanai M, Shiozaki D, Sakimoto S, et al. Association of disorganization of retinal inner layers with optical coherence tomography angiography features in branch retinal vein occlusion. *Graefes Arch Clin Exp Ophthalmol*. 2021 Apr 16. https://doi.org/10.1007/s00417-021-05168-2. Online ahead of print.
29. Schmidt-Erfurth U, Michl M. Disorganization of retinal inner layers and the importance of setting boundaries. *JAMA Ophthalmol*. 2019;137(1):46−47.
30. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull*. 1971;76(05):378−382.
31. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159−174.
32. Cohen Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20(1):37−46.
33. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*. 2012;22(3):276−282.
34. Muftuoglu IK, Mendoza N, Gaber R, et al. Integrity of outer retinal layers after resolution of central involved diabetic macular edema. *Retina*. 2017;37(11):2015−2024.
35. Sakai T, Okude S, Tsuneoka H. Foveal threshold and photoreceptor integrity for prediction of visual acuity after intravitreal aflibercept on age-related macular degeneration. *Clin Ophthalmol*. 2018;12:719−725.
36. Itoh Y, Vasanji A, Ehlers JP. Volumetric ellipsoid zone mapping for enhanced visualisation of outer retinal integrity with optical coherence tomography. *Br J Ophthalmol*. 2016;100(3):295−299.
37. Etheridge T, Dobson ETA, Wiedenmann M, et al. A semi-automated machine-learning based workflow for ellipsoid zone analysis in eyes with macular edema: SCORE2 pilot study. *PLoS One*. 2020;15(4):e0232494.
38. Wang Z, Camino A, Hagag AM, et al. Automated detection of preserved photoreceptor on optical coherence tomography in choroideremia based on machine learning. *J Biophotonics*. 2018;11(5):e201700313.