ARTICLE

Check for updates

A systematic evaluation of human expert agreement on optical coherence tomography biomarkers using multiple devices

Martin Michl[1,6], Martina Neschi[2,6], Alexandra Kaider[3], Katja Hatz[1,4,5], Gabor Deak[1], Bianca S. Gerendas[1]✉ and Ursula Schmidt-Erfurth[1]

OBJECTIVES: To assess the agreement in evaluating optical coherence tomography (OCT) variables in the leading macular diseases such as neovascular age-related macular degeneration (nAMD), diabetic macular oedema (DMO) and retinal vein occlusion (RVO) among OCT-certified graders.
METHODS: SD-OCT volume scans of 356 eyes were graded by seven graders. The grading included presence of intra- and subretinal fluid (IRF, SRF), pigment epithelial detachment (PED), epiretinal membrane (ERM), conditions of the vitreomacular interface (VMI), central retinal thickness (CRT) at the foveal centre-point (CP) and central millimetre (CMM), as well as height and location of IRF/SRF/PED. Kappa statistics (κ) and intraclass correlation coefficient (ICC) were used to report categorical grading and measurement agreement.
RESULTS: The overall agreement on the presence of IRF/SRF/PED was κ = 0.82/0.85/0.81; κ of VMI condition was 0.77, that of ERM presence 0.37. ICC for CRT measurements at CP and CMM was excellent with an ICC of 1.00. Height measurements of IRF/SRF/PED showed robust consistency with ICC = 0.85–0.93. There was substantial to almost perfect agreement in locating IRF/SRF/PED with κ = 0.67–0.86. Between diseases, κ of IRF/SRF presence was 0.69/0.80 for nAMD, 0.64/0.83 for DMO and 0.86/0.89 for RVO.
CONCLUSION: Even in the optimized setting, featuring certified graders, standardized image acquisition and the use of a professional reading platform, there is a disease dependent variability in biomarker evaluation that is most pronounced for IRF in nAMD as well as DMO. Our findings highlight the variability in the performance of human expert OCT grading and the need for AI-based automated feature analyses.

Eye (2023) 37:2573–2579; https://doi.org/10.1038/s41433-022-02376-w

INTRODUCTION

Accurate identification of optical coherence tomography (OCT) biomarkers is essential for an adequate performance of anti-vascular endothelial growth factor (anti-VEGF) therapy in exudative macular diseases, a task that is not only time-intensive, but subjective and prone to error [1]. Reading centres (RC) have therefore become important players in the conduct of clinical trials, not only for patient eligibility, but also for standardized data acquisition and independent image grading [2–4]. As defined in the standard operating procedures, RC graders are trained, certified and continuously supervised by experienced graders or clinicians, thereby achieving a high degree of standardization.

Particularly fluid-related features have led to controversial discussions in human expert assessments. The FLUID study tested the hypothesis that residual subretinal fluid (SRF) in a flexible treatment regimen in patients with neovascular age-related macular degeneration (nAMD) does not entail inferior visual outcomes than when all SRF is resolved. Spectral-domain SD-OCT images in that study were assessed by both RC and on-site investigators, leading to noticeable disagreements in the assessment of intraretinal fluid (IRF) and SRF [5]. Furthermore, post-hoc analyses of the FLUID study using artificial intelligence-based precision tools demonstrated that there was no quantitative difference in SRF between SRF-tolerant and SRF-intolerant treat-and-extend regimens [6].

Even among retina specialists, there is significant disagreement in identifying patients with retinal fluid or referable retinal disease, when assessed by OCT [7, 8]. In an analysis of the Comparison of Age-Related Macular Degeneration Treatment Trials (CATT), both clinicians and RC personnel assessed over 6000 OCT scans for the presence of macular fluid, resulting in a disagreement of 27.9% [1]. These examples highlight a considerable variability in OCT image assessment, with potentially sight-threatening implications for the patients.

Although the human performance of OCT grading has been subject to numerous studies in the past, there are distinct characteristics to all of them, limiting a meaningful comparison of their outcomes. While there is a larger body of such studies focusing on nAMD [2, 9–14], there is little literature on the reproducibility of assessing OCT changes in diabetic macular oedema (DMO) [15–17] or macular oedema due to retinal vein

[1]Department of Ophthalmology, Vienna Reading Center, Medical University of Vienna, Vienna, Austria. [2]RetInSight, Vienna, Austria. [3]Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Vienna, Austria. [4]Faculty of Medicine, University of Basel, Basel, Switzerland. [5]Vista Augenklinik Binningen, Binningen, Switzerland. [6]These authors contributed equally: Martin Michl, Martina Neschi. ✉email: bianca.gerendas@meduniwien.ac.at

Received: 14 October 2022 Revised: 2 December 2022 Accepted: 16 December 2022
Published online: 28 December 2022

Table 1. Descriptive characteristics of the study cohort. 70% of eyes had received three or fewer anti-VEGF injections.

	nAMD	DMO	RVO	Total cohort
Patients, *n* (%)	69 (33.3)	69 (33.3)	69 (33.3)	207
Eyes, *n* (%)	97 (27.25)	138 (38.76)	121 (33.99)	356
≤3 injections, n (%)	*70 (72.16)*	*96 (69.57)*	*84 (69.42)*	250 *(70.2)*
>3 injections, n (%)	*27 (27.84)*	*42 (30.43)*	*37 (30.58)*	106 *(29.8)*
Cirrus scans, *n* (%)	44 (26.04)	66 (39.05)	59 (34.91)	169
Spectralis scans, *n* (%)	48 (28.07)	66 (38.6)	57 (33.33)	171
Topcon scans, *n* (%)	5 (31.25)	6 (37.5)	5 (31.25)	16

n number, *nAMD* neovascular age-related macular degeneration, *DMO* diabetic macular oedema, *RVO* retinal vein occlusion.
Italic values denote ≤3 injections eyes treated with three or fewer injections, >3 injections eyes treated with more than three injections.

occlusion (RVO) [18–20]. With many of these studies dating back around a decade, the most widely used OCT was time-domain TD-OCT. More recent studies compared TD- with SD-OCT [10, 18] or applied SD- or swept source OCT only [9, 15, 17]. Importantly, there is no study that included all three retinal diseases and assessed OCT grading agreement in the most commonly used SD-OCT imaging devices.

Due to rapid advances of artificial intelligence (AI) in ophthalmology, the manual and laborious image assessment is about to be drastically changed and replaced by objective and automated imaging tools [21]. A prerequisite for the implementation of such intelligent tools into clinical practice is not only to prove their non-inferiority to conventional methods, but to identify the limits of a manual assessment and thereby establish benchmarks for human diagnostic performance [22].

The aim of this study is to systematically determine the agreement among OCT-certified graders in assessing the key OCT features that are relevant for the morphological assessment of macular diseases and are routinely evaluated both in a clinical routine and trial setting.

METHODS

Population/dataset

In this post-hoc analysis, we included SD-OCT imaging data of five randomized multi-centre clinical trials from the Vienna Reading Center (VRC) imaging database. Patients were affected by nAMD, DMO or branch/central RVO (BRVO/CRVO) and were equally represented in the dataset (Table 1). In relation to the number of scans recorded with each OCT device in the clinical trials, we randomly selected scans taken with Cirrus HD-OCT (Carl Zeiss Meditec, Dublin, CA, USA, software version 4.5 or later), Spectralis OCT or HRA-OCT (Heidelberg Engineering, Heidelberg, Germany, software version 4.0.0.0 or later) and Topcon 3D OCT-1000 or 3D OCT-2000 (Topcon Corp., Tokyo, Japan) (Table 1). While all devices covered a 6 ×6 mm area, Cirrus and Topcon volumes comprised 128 b-scans and Spectralis volumes 49 b-scans; RC fovea-centred scans were used (Fig. 1). Baseline to month 3 visits (= patients had received a maximum of three anti-VEGF injections) comprised 70% of all eyes to increase the likelihood of multiple OCT feature presence.

All patients had to give their informed consent before entering the respective studies and ethical approval was obtained from each participating centre's institutional review board for inclusion in the trials and consecutive scientific analyses. All patient-identifying data were removed from image data. The analysis of the data adhered to the tenets of the Declaration of Helsinki and was approved by the ethics committee of the Medical University of Vienna (approval number: 1246/2016).

Grading process

Seven OCT-certified and masked graders of the VRC independently examined all images following a predefined reading protocol in a custom VRC software. All graders had grading experience on exudative macular diseases and had received formal OCT training exclusively on nAMD eyes. Since the goal of this study was to compare the grading performance of the individual graders, there was no additional supervision during the grading process. Our graders received no specific harmonization training for this study to avoid introducing bias and to identify OCT variables that are most susceptible to reproducibility issues under such conditions.

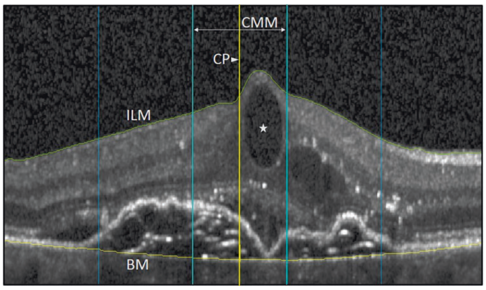

Fig. 1 An ETDRS grid was centred on the foveal centre-point (CP) by the reading centre for comparability of feature localization among all graders. Manual correction of the inner limiting membrane (ILM) and Bruch's membrane (BM) were performed to generate retinal thickness measurements at the CP and the central one millimetre subfield (CMM). As seen here, retinal thickness thus included subretinal fluid and a pigment epithelial detachment, if present. The star marks intraretinal cystoid fluid.

The graders manually corrected the segmentation lines of the internal limiting membrane (ILM) and Bruch's membrane (BM) to acquire retinal thickness measurements at the CP and CMM. The latter thus included subretinal fluid (SRF) and pigment epithelial detachment (PED), if present. It has been shown before that the line correction in our custom software leads to reliable correlations of thickness measurements across all included OCT devices [23].

The morphological grading comprised the following OCT parameters: presence/absence and location (CP; CMM; outside CMM) of IRF, SRF, PED, macular hole (MH), macular atrophy (MA); visibility of the vitreomacular interface (VMI) and the condition thereof (e.g., vitreomacular adhesion, VMA = adhesion is visible in >90% of scans; partial vitreomacular adhesion, PVMA = adhesion visible in ≤90%; vitreomacular traction, VMT; full posterior vitreous detachment, PVMA) and epiretinal membrane (ERM). If present at the CP, the height of IRF, SRF and PED was measured perpendicular to Bruch's membrane.

Statistical analysis

The calculation of inter- and intragrader agreement for a specific feature was based on images where all graders agreed that image quality was adequate for assessment.

Intergrader agreement for qualitative measures was determined by means of the generalized Cohen's kappa coefficient for multiple raters and its 95% confidence intervals (CI): presence/absence and the location of IRF, SRF, PED, MH, MA, ERM and condition of the VMI. A kappa coefficient between 0.01–0.20 was described as slight agreement; 0.21–0.40=fair; 0.41–0.60=moderate; 0.61–0.80=substantial; 0.81–1.00=almost perfect agreement [24]. The agreement of retinal thickness measurements and feature heights was described with the intraclass correlation coefficient (ICC): less than 0.50=poor agreement; 0.50–0.75=moderate; 0.75–0.90=good; above 0.90=excellent agreement [25]. Intergrader agreement for ordinal data (degree of posterior vitreous detachment, PVD) was assessed using Kendall's Coefficient of Concordance (Kendall's W): 0.5–0.7=moderate agreement; 0.7–0.9=strong agreement; 0.9–1=unusually strong agreement.

We further examined the intergrader agreement separately for each OCT device (Spectralis, Cirrus, Topcon), the three diseases (nAMD, DMO, RVO) and treatment stage (≤month 3; >3 months).

For the assessment of *intragrader agreement*, graders re-evaluated 10% of all images. The latter were randomly selected, re-numbered, and the graders were not informed about the inclusion of duplicates in the dataset. Due to their skew distributions, log2-transformed values of the continuous variables (retinal thickness measurements and feature heights) were used for statistical analysis. OCT features present in less than 5% of all scans were excluded from the analysis. Statistical analyses were performed using the software SAS 9.4 (SAS Institute Inc., 2016. Cary, NC, USA).

RESULTS

Three hundred fifty-six eyes (49.1% left eyes) from 207 patients were selected for inter- and intragrader agreement assessment (Table 1): 97 eyes with nAMD, 138 eyes with DMO and 121 eyes with BRVO/CRVO. The randomly selected scans comprised 169 Cirrus scans (47,5%), 171 Spectralis scans (48%) and 16 Topcon scans (4,5%).

Intergrader agreement

1. *Feature presence/location (*Table 2A, B*)*. There was almost perfect agreement in the assessment of overall IRF, SRF and PED presence, substantial agreement for VMI, and fair agreement for the grading of ERM. There was substantial to almost perfect agreement in locating IRF, SRF and PED at the CP, CMM and outside the CMM (kappa: 0.67–0.86).
2. *Retinal thickness/feature height (*Table 2A*)*. Segmentation line corrections of the ILM and BM resulted in excellent agreement between all graders for retinal thickness values both at the CP and CMM (both ICC: 1.0). The agreement of grading IRF-, SRF-, and PED height at the CP was comparable with an ICC of 0.92 (excellent), 0.85 (good) and 0.93 (excellent), respectively.
3. *Vitreomacular interface status*. There was fair agreement between graders in assessing conditions of the VMI (kappa: 0.25; CI: 0.18–0.31), but moderate agreement in assessing the degree of PVD (Kendall's W: 0.58).
4. *Comparison between OCT devices (*Table 2C*)*. Slightly higher intergrader agreement was observed in assessing the presence of IRF, PED, VMI, ERM and IRF/SRF height in Spectralis scans, compared to Cirrus scans. There was almost no difference between devices for retinal thickness measurements and PED height. Topcon scans were underrepresented (<5%) and are thus not part of this sub-analysis.
5. *Comparison between treatment stages (*Table 2C*)*. There were no major differences in grading the OCT parameters before or after the completion of the anti-VEGF loading phase (= before or after month 3).
6. *Comparison between diseases (*Table 3*)*. There were no differences in assessing retinal thickness, but moderate (nAMD), good (RVO) and excellent (DMO) agreement in assessing SRF height.

IRF presence was graded with substantial agreement in nAMD and DMO, and almost perfect agreement in RVO. While in nAMD no macular hole was identified, less than 5% of eyes with DMO/RVO were affected by this feature. An incidence of more than 5% for MA was seen in nAMD eyes (6.9%) and DMO eyes (5.4%) with moderate and slight grading agreement between graders, respectively.

Figure 2 shows example images that led to obvious disagreements between the graders.

Intragrader agreement

Intragrader agreement was calculated based on the re-grading of 37 eyes by the same expert. All graders showed excellent agreement in assessing retinal thickness and feature heights (ICC: 1.00). While the presence of IRF, PED and VMI conditions was graded with substantial up to almost perfect agreement (kappa: 0.77–1.0), the presence of SRF and ERM was graded ‚reproducibly' with kappa values between 0.70 and 0.89. Two graders had moderate intragrader agreement when grading the presence of SRF (kappa: 0.56) and one grader re-graded the presence of ERM with fair agreement (kappa: 0.28).

DISCUSSION

The goal of this study was to systematically assess the agreement of individual human grading experts in evaluating both qualitative as well as quantitative OCT markers in a large representative set of eyes affected by nAMD, DMO or macular oedema due to RVO, a task that is routinely performed in clinical routine and trial settings. As the presence of retinal fluid impacts functional loss as well as treatment decisions in all three diseases, they were the primary focus of this study. Although IRF and SRF might appear similar in nAMD, DMO and RVO, there is no study comparing their actual gradability between the three conditions: interestingly, our results show an obvious difference in the agreement on IRF presence in nAMD (kappa 0.69), DMO (0.64) and RVO (0.86). Such levels of disagreement among certified human experts are surprising, especially in an optimized setting that features standardized image acquisition following defined study protocols and a user-friendly platform for professional image grading.

In RVO, the higher relative agreements on IRF presence correspond to previously published results in TD-OCT of 76%–83% [19] and 84% [20]. Typically, large cystoid spaces in a substantially thickened retina affect the ganglion cell- as well as inner- and outer nuclear layer and are seen sub- or parafoveally, making them overall more easily identifiable [26]. This is in contrast to the often subtle hyporeflective spaces in nAMD that might be confused with "pixel voids", a term that was described as a cyst-like appearance of hyporeflectivity in hyporeflective retinal layers due to low signal intensity, in the absence of actual cystoid changes [1, 2]. The reduced consistency in grading IRF in nAMD might be due to the association of numerous other structural alterations, many of them being degenerative in nature, whereas in RVO, IRF occurs as an acute spreading of fluid into an otherwise unaltered retina. Depending on the underlying lesion type in nAMD, there is additional variability in retinal fluid localisation and extension [27]. In a post-hoc analysis of 270 TD-OCT scans from the CATT, DeCroos et al. assessed the reproducibility between two independent grading teams and found kappas of 0.48, 0.8 and 0.75 for the detection of IRF, SRF and sub-RPE fluid, respectively [2]. In another retrospective case series with AMD eyes, four independently trained retina specialists were asked to grade 112 SD-OCT images and reached agreements of 0.62, 0.82 and 0.60 for the detection of IRF, SRF and PED, respectively, producing similar results as those in our AMD cohort [9]. When comparing the detection of macular fluid between ophthalmologists and a RC, major causes of disagreements were found to be thinner retinas, smaller fluid pockets and greater decrease of retinal thickness at the foveal centre [1]. Keenan et al. confirmed these findings in a recent follow-on study of the AREDS 2 trial. It compared the performance of retina specialists in assessing retinal fluid in SD-OCT images to a deep learning-based algorithm and reported an accuracy of 0.81, a sensitivity of 0.47 and a specificity of 0.97. It was found that IRF was significantly more often missed by the graders when appearing in the absence of SRF or if the mean retinal fluid volume and number of b-scans showing fluid was lower [7]. One may assume that the same factors might also complicate the grading of IRF in DMO [28]. Albeit comparable to nAMD, a kappa of 0.64 for the detection of IRF in DMO was nevertheless surprising. Worse reproducibility might also be due

Table 2. Intergrader agreement on (A) OCT feature presence as well as retinal thickness/feature height for the whole study cohort (all diseases), (B) feature localization at the centre point (CP), within the central millimetre (CMM) and outside the CMM, and (C) on feature presence and retinal thickness/feature height with regard to OCT device and treatment stage in the total patient cohort.

A. Feature presence	**Kappa (CI)**	***n***	**Agreement**
IRF	0.82 (0.78–0.86)	352	almost perfect
SRF	0.85 (0.80–0.89)	349	almost perfect
PED	0.81 (0.76–0.85)	350	almost perfect
VMI[a]	0.77 (0.73–0.80)	354	substantial
ERM[b]	0.37 (0.32–0.41)	354	fair
MH	°	°	°
MA	°	°	°
Retinal thickness/feature height	**ICC**	***n***	**Agreement**
Thickness at CP	1.00	354	excellent
Thickness at CMM	1.00	354	excellent
IRF height[a]	0.92	58	excellent
SRF height[b]	0.85	39	good
PED height[b]	0.93	36	excellent

B. Feature location	**CP**		**CMM**		**Outside CMM**	
	Kappa (CI)	***n***	**Kappa (CI)**	***n***	**Kappa (CI)**	***n***
IRF	0.78 (0.73–0.83)	352	0.83 (0.79–0.86)	352	0.81 (0.77–0.84)	352
SRF	0.86 (0.81–0.91)	349	0.85 (0.81–0.89)	349	0.84 (0.80–0.89)	349
PED	0.67 (0.61–0.74)	350	0.77 (0.71–0.82)	350	0.80 (0.75–0.85)	350

C. Feature presence	**OCT device**				**Treatment stage**			
	Cirrus		**Spectralis**		**≤M3**		**>M3**	
	Kappa (CI)	***n***	**Kappa (CI)**	***n***	**Kappa (CI)**	***n***	**Kappa (CI)**	***n***
IRF	0.78 (0.72–0.83)	167	0.85 (0.80–0.90)	170	0.81 (0.77–0.86)	249	0.83 (0.77–0.90)	103
SRF	0.86 (0.80–0.91)	166	0.83 (0.76–0.90)	168	0.86 (0.82–0.90)	247	0.77 (0.65–0.90)	102
PED	0.79 (0.72–0.87)	167	0.80 (0.74–0.87)	168	0.80 (0.74–0.86)	248	0.82 (0.74–0.90)	102
VMI[a]	0.73 (0.67–0.78)	168	0.83 (0.77–0.88)	171	0.74 (0.69–0.78)	250	0.84 (0.79–0.90)	104
ERM[b]	0.29 (0.22–0.35)	168	0.42 (0.34–0.49)	171	0.34 (0.29–0.40)	250	0.41 (0.32–0.50)	104
MH	°	°	°	°	°	°	°	°
MA	°	°	°	°	°	°	°	°
Retinal thickness/feature height	**ICC**	***n***	**ICC**	***n***	**ICC**	***n***	**ICC**	***n***
Thickness at CP	1.00	168	1.00	171	1.00	250	1.00	104
Thickness at CMM	1.00	168	1.00	171	1.00	250	1.00	104
IRF height[a]	0.89	24	0.95	27	0.91	44	0.98	14
SRF height[b]	0.85	27	0.88	10	0.84	34	0.93	5
PED height[b]	0.94	13	0.94	22	0.92	24	0.94	12

IRF intraretinal fluid, *SRF* subretinal fluid, *PED* pigment epithelial detachment, *VMI* vitreomacular interface, *ERM* epiretinal membrane, *MH* macular hole, *MA* macular atrophy, *CP* centre point, *CMM* central millimetre, *CI* confidence interval, *n* number of observations, *ICC* intraclass correlation coefficient, *≤M3* before completion of anti-VEGF loading phase, *>M3* after completion of anti-VEGF loading phase.

° feature found in less than 5% of images.

[a]one, [b]two graders indicated 'not received training for this feature'.

to small focal oedemas off the macular centre that are more easily missed, especially when overshadowed by hard exudates.

In contrast to the difficult task of IRF assessment, more consistent grading results in our study were found for SRF (nAMD: 0.8; DMO: 0.83; RVO: 0.89). This is not unexpected for a feature that affects an anatomically predefined space of the retina and therefore shows less variability in appearance. In nAMD, however, the association of (heterogeneous) hyperreflective material in the subretinal compartment as well as outer retinal degeneration might complicate the grading of SRF. Nonetheless, various studies produced comparable results for the detection of SRF in nAMD with kappas ranging from 0.72 to 0.82 [2, 9, 10, 12]. While less comparable results were found in a small-scale DMO study [28], there is so far no literature on OCT grading agreement of SRF in RVO.

As seen in previous studies [2, 10–12, 15, 19], measuring CST reached excellent agreements between our graders (ICC:1.0), independent of the disease. While CST is still used as an anatomical outcome measure in clinical trials, it is neither a reliable indicator of disease activity nor does it show a meaningful correlation with visual function over time [29–33]. Pawloff et al.

Table 3. Intergrader agreement on (A) feature presence and (B) retinal thickness/feature height, separately for each disease.

A. Feature presence	**nAMD**		**DMO**		**RVO**	
	Kappa (CI)	***n***	**Kappa (CI)**	***n***	**Kappa (CI)**	***n***
IRF	0.69 (0.60–0.78)	95	0.64 (0.48–0.81)	138	0.86 (0.81–0.91)	119
SRF	0.80 (0.72–0.87)	95	0.83 (0.74–0.91)	135	0.89 (0.82–0.95)	119
PED	0.56 (0.45–0.67)	95	°		°	
VMI[a]	0.78 (0.71–0.85)	95	0.79 (0.73–0.85)	138	0.73 (0.67–0.79)	121
ERM[b]	0.31 (0.21–0.40)	95	0.34 (0.27–0.41)	138	0.42 (0.33–0.51)	121
MH	no subject	-	°	°	°	°
MA[b]	0.50 (0.30-0.69)	95	0.19 (0.10-0.27)	137	°	°
B. Retinal thickness/feature height	**ICC**	***n***	**ICC**	***n***	**ICC**	***n***
Thickness at CP	1.00	95	1.00	138	1.00	121
Thickness at CMM	1.00	95	1.00	138	1.00	121
IRF height[a]	0.98	9	0.87	37	0.96	12
SRF height[b]	0.73	14	0.97	10	0.89	15
PED height[b]	0.93	36	°	°	no subject	-

IRF intraretinal fluid, *SRF* subretinal fluid, *PED* pigment epithelial detachment, *VMI* vitreomacular interface, *ERM* epiretinal membrane, *MH* macular hole, *MA* macular atrophy, *CP* centre point, *CMM* central millimetre, *nAMD* neovascular age-related macular degeneration, *DMO* diabetic macular oedema, *RVO* retinal vein occlusion, *CI* confidence interval, *n* number of observations, *ICC* intraclass correlation coefficient.

° feature found in less than 5% of images.

[a]one, [b]two graders indicated 'not received training for this feature'.

applied a precision AI fluid algorithm to more than 2400 eyes to assess the correlation of (three-dimensional) retinal fluid volumes and (two-dimensional) central retinal subfield thickness, demonstrating a surprisingly low correlation of $r = 0.57$ in nAMD [32].

With the increasingly important role of retinal imaging in clinical trials, there has been a growing trend towards centralizing decisions on patient eligibility and disease monitoring in RCs. To ensure a high degree of standardization while keeping bias and variability at a minimum, RCs operate under standards that cover not only technical aspects, but also image acquisition, interpretation and documentation. Image gradings are based on clear feature definitions and are performed by certified graders who receive a study- and/or disease specific training. Gradings are often based on a dual reading, where images are assessed by two independent graders who are supervised by a third more experienced grader or retina specialist. In our study, no additional training nor supervision was conducted. Therefore, the results presented herein do not fully reflect the general reproducibility of OCT grading at a RC, but more importantly the reproducibility of a human expert grading in the real-world.

While the employment of any of these RC-specific measures in a real-world practice might be beneficial for patient and clinician, their adoption will likely be complicated by cost and time constraints, as well as the professional training and experience of the individual clinician: CATT was an important endeavour that compared the treatment decision by ophthalmologists versus that of a RC [1]. Any macular fluid, as seen on OCT, mandated the administration of anti-VEGF injections. Prior to study initiation, treating ophthalmologists were required to perform an investigator training and pass a knowledge assessment test involving the interpretation of OCTs. Notwithstanding, there were marked discrepancies in the identification of macular fluid in 1737 of 6210 visits (=28%), most commonly in visits where the RC detected macular fluid while clinicians did not. This is of significant relevance, considering that in nAMD, ophthalmologists prefer to base their treatment decisions on structural OCT changes rather than visual acuity or FDA labelling (American Society of Retina Specialists, 2020. Global Trends in Retina).

The limited reproducibility seen in our and previous studies raises the question whether OCT image assessment, as we know it, has reached its maximum potential. Despite the resources available to a RC, manual gradings remain laborious, and to a certain extent inconsistent and inefficient: OCT imaging holds information that is generated by millions of pixels per volume; however, an image grading that merely assesses qualitative aspects (e.g., feature presence/absence) or two-dimensional parameters (e.g., CST, feature height) does not capture the great quantity of available structural data.

Artificial intelligence (AI)-based algorithms are promising tools that allow a more precise and objective evaluation of the continuously increasing imaging data. An automated detection of retinal fluid is capable of determining not only fluid presence, but also subtype, location and volume [21]. An accurate assessment of retinal fluid in exudative diseases is most important, as increasing fluid volumes at each compartment have been shown to negatively impact BCVA outcomes, independent of the therapeutic substance used [34]. Recording compartmental and volume-based parameters over time will help to identify clinically meaningful thresholds for retinal fluid and standardize treatment decisions between clinicians. This is vital for the individual patient, as both under- and overtreatment should be avoided at any cost. While the translation of findings from clinical trials to the general population is often limited by strict inclusion or exclusion criteria, the application of objective metrics in this setting might mitigate this problem. Most importantly, because the results of automated feature detection can be shared by the cloud in real time, study sites could be freed from the delayed feedback of RCs, thereby enhancing patient enrolment and study visits. Real-time AI-based feedback at any level of a randomized clinical trial (e.g., patient screening, monitoring and final data analysis) substantially saves human and financial resources and increases the transparency for investigators and sponsors.

While the focus of this study was on exudative changes, the sample size was too small to come to conclusions on less frequently seen OCT changes such as epiretinal membranes, macular holes and macular atrophy. Graders are typically aware of the underlying condition when grading; it is uncertain whether the simultaneous presentation of OCT images from different diseases introduced a grading bias (e.g., a small area of hyporeflectivity in an eye with nAMD might more readily be graded as IRF if presented after a consecutive series of eyes with RVO showing obvious IRF).

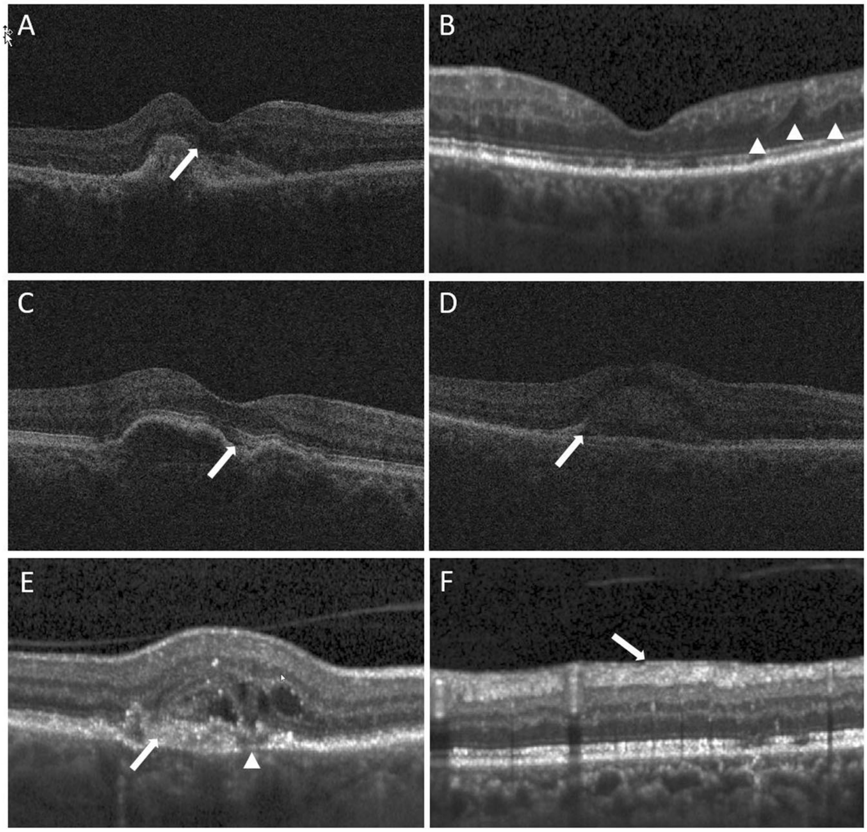

Fig. 2 Example images of cases leading to disagreement between graders. A The identification of small quantities of IRF (arrow) at the central subfield is more difficult due to the absence of inner retinal layers, especially in low contrast images. **B** Irregularities at the level of the outer plexiform/nuclear layer and the presence of hyperreflective foci in this RVO eye led to disagreements regarding IRF presence (arrow heads). **C** Graders disagreed on SRF presence (arrow), likely due to its small quantity and its possible mechanical generation between two adjacent RPE detachments. **D** The presence of subretinal hyperreflective material (SHRM) or possibly blood as seen in this eye with nAMD might have led to the disagreement on the presence of SRF (arrow). **E** The identification of a pigment epithelial detachment in nAMD led to moderate agreement between graders. Disagreement was often due to degenerative changes of the outer retina, such as (partial) loss of the retinal pigment epithelium (arrow head), fibrosis as well as SHRM (arrow). **F** The grading of an epiretinal membrane led to fair agreement between graders. Disagreement was often due to subtle hyperreflective lines with no associated irregularities of the underlying retinal layers (arrow).

In conclusion, our systematic evaluation of human expert agreement on OCT biomarkers in nAMD, DMO and RVO found an SRF agreement that was rather consistent across all three conditions. However, there was a substantial grading disagreement concerning IRF in nAMD and DMO. Importantly, any image assessment by a human, even in the highly standardized setting of a RC, remains laborious and to a certain degree subjective. Our goal should therefore focus on the adoption of automated imaging analysis tools for a more precise, efficient and objective image assessment. Furthermore, enhanced collaborations of different reading centres in large-scale clinical studies call for the harmonization and standardization of grading procedures not only within, but between centres.

Summary

What was known before

- Previous studies have reported on relevant OCT grading (dis-) agreements for individual macular diseases but lack comparability and are mostly based on outdated imaging techniques.

What this study adds

- This reading-centre-based study is the first to compare SD-OCT biomarker grading in the most prevalent exudative macular diseases.
- The observed (dis-) agreements depend on the underlying disease and are most striking for retinal fluid.
- The consistency of OCT grading by human experts is limited, even in the most standardized setting.

DATA AVAILABILITY

Data are available upon reasonable request. Data may be obtained from a third party and are not publicly available.

REFERENCES

1. Toth CA, Decroos FC, Ying GS, Stinnett SS, Heydary CS, Burns R, et al. Identification of Fluid on Optical Coherence Tomography by Treating Ophthalmologists Versus a Reading Center in the Comparison of Age-Related Macular Degeneration Treatments Trials. Retina. 2015;35:1303–14.
2. DeCroos FC, Toth CA, Stinnett SS, Heydary CS, Burns R, Jaffe GJ, et al. Optical coherence tomography grading reproducibility during the Comparison of Age-related Macular Degeneration Treatments Trials. Ophthalmology. 2012;119:2549–57.
3. Mitchell P, Bandello F, Schmidt-Erfurth U, Lang GE, Massin P, Schlingemann RO, et al. The RESTORE study: ranibizumab monotherapy or combined with laser versus laser monotherapy for diabetic macular edema. Ophthalmology. 2011;118:615–25.
4. Nguyen QD, Brown DM, Marcus DM, Boyer DS, Patel S, Feiner L, et al. Ranibizumab for diabetic macular edema: results from 2 phase III randomized trials: RISE and RIDE. Ophthalmology. 2012;119:789–801.

5. Guymer RH, Markey CM, McAllister IL, Gillies MC, Hunyor AP, Arnold JJ, et al. Tolerating Subretinal Fluid in Neovascular Age-Related Macular Degeneration Treated with Ranibizumab Using a Treat-and-Extend Regimen: FLUID Study 24-Month Results. Ophthalmology. 2019;126:723–34.
6. Reiter GS, Grechenig C, Vogl WD, Guymer RH, Arnold JJ, Bogunovic H, et al. Analysis of fluid volume and its impact on visual acuity in the FLUID study as quantified with deep learning. Retina. 2021;41:1318–28.
7. Keenan TDL, Clemons TE, Domalpally A, Elman MJ, Havilio M, Agron E, et al. Retinal Specialist versus Artificial Intelligence Detection of Retinal Fluid from OCT: Age-Related Eye Disease Study 2: 10-Year Follow-On Study. Ophthalmology. 2021;128:100–9.
8. De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med. 2018;24:1342–50.
9. Muller PL, Liefers B, Treis T, Rodrigues FG, Olvera-Barrios A, Paul B, et al. Reliability of Retinal Pathology Quantification in Age-Related Macular Degeneration: Implications for Clinical Trials and Machine Learning Applications. Transl Vis Sci Technol. 2021;10:4.
10. Folgar FA, Jaffe GJ, Ying GS, Maguire MG, Toth CA, Comparison of Age-Related Macular Degeneration Treatments Trials Research G. Comparison of optical coherence tomography assessments in the comparison of age-related macular degeneration treatments trials. Ophthalmology. 2014;121:1956–65.
11. Joeres S, Tsong JW, Updike PG, Collins AT, Dustin L, Walsh AC, et al. Reproducibility of quantitative optical coherence tomography subanalysis in neovascular age-related macular degeneration. Invest Ophthalmol Vis Sci. 2007;48:4300–7.
12. Ritter M, Elledge J, Simader C, Deak GG, Benesch T, Blodi BA, et al. Evaluation of optical coherence tomography findings in age-related macular degeneration: a reproducibility study of two independent reading centres. Br J Ophthalmol. 2011;95:381–5.
13. Zhang N, Hoffmeyer GC, Young ES, Burns RE, Winter KP, Stinnett SS, et al. Optical coherence tomography reader agreement in neovascular age-related macular degeneration. Am J Ophthalmol. 2007;144:37–44.
14. Sayegh RG, Simader C, Scheschy U, Montuoro A, Kiss C, Sacu S, et al. A systematic comparison of spectral-domain optical coherence tomography and fundus autofluorescence in patients with geographic atrophy. Ophthalmology. 2011;118:1844–51.
15. Sala-Puigdollers A, Figueras-Roca M, Hereu M, Hernandez T, Morato M, Adan A, et al. Repeatability and reproducibility of retinal and choroidal thickness measurements in Diabetic Macular Edema using Swept-source Optical Coherence Tomography. PLoS One. 2018;13:e0200819.
16. Glassman AR, Beck RW, Browning DJ, Danis RP, Kollman C. Diabetic Retinopathy Clinical Research Network Study G. Comparison of optical coherence tomography in diabetic macular edema, with and without reading center manual grading from a clinical trials perspective. Invest Ophthalmol Vis Sci. 2009;50:560–6.
17. Munk MR, Lincke J, Giannakaki-Zimmermann H, Ebneter A, Wolf S, Zinkernagel MS. Comparison of 55 degrees Wide-Field Spectral Domain Optical Coherence Tomography and Conventional 30 degrees Optical Coherence Tomography for the Assessment of Diabetic Macular Edema. Ophthalmologica. 2017;237:145–52.
18. Hatef E, Khwaja A, Rentiya Z, Ibrahim M, Shulman M, Turkcuoglu P, et al. Comparison of time domain and spectral domain optical coherence tomography in measurement of macular thickness in macular edema secondary to diabetic retinopathy and retinal vein occlusion. J Ophthalmol. 2012;2012:354783.
19. Domalpally A, Blodi BA, Scott IU, Ip MS, Oden NL, Lauer AK, et al. The Standard Care vs Corticosteroid for Retinal Vein Occlusion (SCORE) study system for evaluation of optical coherence tomograms: SCORE study report 4. Arch Ophthalmol. 2009;127:1461–7.
20. Decroos FC, Stinnett SS, Heydary CS, Burns RE, Jaffe GJ. Reading Center Characterization of Central Retinal Vein Occlusion Using Optical Coherence Tomography During the COPERNICUS Trial. Transl Vis Sci Technol. 2013;2:7.
21. Schmidt-Erfurth U, Reiter GS, Riedl S, Seebock P, Vogl WD, Blodi BA, et al. AI-based monitoring of retinal fluid in disease activity and under therapy. Prog Retin Eye Res. 2021;86:100972
22. Jill Hopkins J, Keane PA, Balaskas K. Delivering personalized medicine in retinal care: from artificial intelligence algorithms to clinical application. Curr Opin Ophthalmol. 2020;31:329–36.
23. Simader C, Montuoro A, Waldstein S, Gerendas B, Lammer J, Heiling U, et al. Retinal Thickness Measurements with Spectral Domain Optical Coherence Devices from Different Manufacturers in a Reading Center Environment. Investigative Ophthalmol Vis Sci. 2012;53:4067.
24. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–74.
25. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016;15:155–63.
26. Michl M, Liu X, Kaider A, Sadeghipour A, Gerendas BS, Schmidt-Erfurth U. The impact of structural optical coherence tomography changes on visual function in retinal vein occlusion. Acta Ophthalmol. 2021;99:418–26.
27. Liakopoulos S, Ongchin S, Bansal A, Msutta S, Walsh AC, Updike PG, et al. Quantitative Optical Coherence Tomography Findings in Various Subtypes of Neovascular Age-Related Macular Degeneration. Investigative Ophthalmol Vis Sci. 2008;49:5048–54.
28. Heng LZ, Pefkianaki M, Hykin P, Patel PJ. Interobserver agreement in detecting spectral-domain optical coherence tomography features of diabetic macular edema. PLoS One. 2015;10:e0126557.
29. Bressler NM, Odia I, Maguire M, Glassman AR, Jampol LM, MacCumber MW, et al. Association Between Change in Visual Acuity and Change in Central Subfield Thickness During Treatment of Diabetic Macular Edema in Participants Randomized to Aflibercept, Bevacizumab, or Ranibizumab: A Post Hoc Analysis of the Protocol T Randomized Clinical Trial. JAMA Ophthalmol. 2019;137:977–85.
30. Jaffe GJ, Martin DF, Toth CA, Daniel E, Maguire MG, Ying GS, et al. Macular morphology and visual acuity in the comparison of age-related macular degeneration treatments trials. Ophthalmology. 2013;120:1860–70.
31. Deák GG, Schmidt-Erfurth UM, Jampol LM. Correlation of Central Retinal Thickness and Visual Acuity in Diabetic Macular Edema. JAMA Ophthalmol. 2018;136:1215–6.
32. Pawloff M, Bogunovic H, Gruber A, Michl M, Riedl S, Schmidt-Erfurth U. Systematic correlation of central subfield thickness with retinal fluid volumes quantified by deep learning in the major exudative macular diseases. Retina. 2022;42:831–41.
33. Gerendas BS, Sadeghipour A, Michl M, Goldbach F, Mylonas G, Gruber A, et al. Validation of an Automated Fluid Algorithm on Real-World Data of Neovascular Age-Related Macular Degeneration over Five Years. Retina. 2022;42:1673–82.
34. Schmidt-Erfurth U, Mulyukov Z, Gerendas BS, Reiter GS, Lorand D, Weissgerber G, et al. Therapeutic response in the HAWK and HARRIER trials using deep learning in retinal fluid volume and compartment analysis. Eye (Lond). (2022). https://doi.org/10.1038/s41433-022-02077-4. Online ahead of print.

ACKNOWLEDGEMENTS

We thank the graders from the Vienna Reading Center for their valuable contribution.

AUTHOR CONTRIBUTIONS

The authors confirm contribution to the paper as follows: study conception and design: MN, GD, BSG, USE; data collection: MN, KH; analysis and interpretation of results: MM, AK, GD, BSG, USE; draft manuscript preparation: MM, BSG, USE; All authors reviewed the results and approved the final version of the manuscript.

COMPETING INTERESTS

BSG: Bayer, Zeiss, Novartis (C), IDx (F); USE: Apellis, Novartis, Kodiak, Roche, Boehringer, RetInSight (C). Other authors declare no competing interests.

ETHICS APPROVAL

Approved by the Ethics Committee of the Medical University of Vienna, #1246/2016.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to Bianca S. Gerendas.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.