tvst

Article

Reliability of Retinal Pathology Quantification in Age-Related Macular Degeneration: Implications for Clinical Trials and Machine Learning Applications

Philipp L. Müller[1–3], Bart Liefers[1,4,5], Tim Treis[6], Filipa Gomes Rodrigues[1,2], Abraham Olvera-Barrios[1,2], Bobby Paul[7], Narendra Dhingra[8], Andrew Lotery[9], Clare Bailey[10], Paul Taylor[11], Clarisa I. Sánchez[4,5,12], and Adnan Tufail[1,2]

[1] Moorfields Eye Hospital NHS Foundation Trust, London, UK
[2] Institute of Ophthalmology, University College London, London, UK
[3] Department of Ophthalmology, University of Bonn, Bonn, Germany
[4] Diagnostic Image Analysis Group, Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Nijmegen, The Netherlands
[5] Department of Ophthalmology, Radboud University Medical Center, Nijmegen, The Netherlands
[6] BioQuant, University of Heidelberg, Heidelberg, Germany
[7] Barking, Havering and Redbridge University Hospitals NHS Trust, Romford, UK
[8] Mid Yorkshire Hospitals NHS Trust, Wakefield, UK
[9] University Hospital Southampton NHS Foundation Trust, Southampton, UK
[10] University Hospitals Bristol NHS Foundation Trust, Bristol, UK
[11] Institute of Health Informatics, University College London, London, UK
[12] Informatics Institute, Faculty of Science, University of Amsterdam, Amsterdam, The Netherlands

Correspondence: Adnan Tufail, Moorfields Eye Hospital NHS Foundation Trust, 162 City Rd, London EC1V 2PD, UK. e-mail: adnan.tufail@nhs.net

Received: October 9, 2020
Accepted: December 22, 2020
Published: March 4, 2021

Keywords: retina; AMD; optical coherence tomography; OCT; imaging; interreader; interrater; agreement; annotation; artificial intelligence; machine learning; deep learning

Citation: Müller PL, Liefers B, Treis T, Rodrigues FG, Olvera-Barrios A, Paul B, Dhingra N, Lotery A, Bailey C, Taylor P, Sánchez CI, Tufail A. Reliability of retinal pathology quantification in age-related macular degeneration: Implications for clinical trials and machine learning applications. Trans Vis Sci Tech. 2021;10(3):4, https://doi.org/10.1167/tvst.10.3.4

Purpose: To investigate the interreader agreement for grading of retinal alterations in age-related macular degeneration (AMD) using a reading center setting.

Methods: In this cross-sectional case series, spectral-domain optical coherence tomography (OCT; Topcon 3D OCT, Tokyo, Japan) scans of 112 eyes of 112 patients with neovascular AMD (56 treatment naive, 56 after three anti–vascular endothelial growth factor injections) were analyzed by four independent readers. Imaging features specific for AMD were annotated using a novel custom-built annotation platform. Dice score, Bland-Altman plots, coefficients of repeatability, coefficients of variation, and intraclass correlation coefficients were assessed.

Results: Loss of ellipsoid zone, pigment epithelium detachment, subretinal fluid, and drusen were the most abundant features in our cohort. Subretinal fluid, intraretinal fluid, hypertransmission, descent of the outer plexiform layer, and pigment epithelium detachment showed highest interreader agreement, while detection and measures of loss of ellipsoid zone and retinal pigment epithelium were more variable. The agreement on the size and location of the respective annotation was more consistent throughout all features.

Conclusions: The interreader agreement depended on the respective OCT-based feature. A selection of reliable features might provide suitable surrogate markers for disease progression and possible treatment effects focusing on different disease stages.

Translational Relevance: This might give opportunities for a more time- and cost-effective patient assessment and improved decision making as well as have implications for clinical trials and training machine learning algorithms.

Introduction

Age-related macular degeneration (AMD) is a leading cause of legal blindness in the industrialized world.[1] Concerning advanced disease manifestations, a dry stage defined by the presence of retinal pigment epithelial (RPE) atrophy (called geographic atrophy [GA]) can be distinguished from or complicated by a neovascular (nAMD) form typically characterized by the presence of choroidal neovascularization (CNV).[2–4]

While both forms of late-stage AMD are associated with the risk of visual loss, an effective treatment for GA development and progression is still pending. However, various therapeutic approaches are tested in different stages of preclinical and clinical trials.[5,6] To accelerate clinical testing, meaningful, validated clinical endpoints are needed.[7] Most interventional trials currently rely on the progression of GA, which is an accepted endpoint by regulators.[8,9] However, the most effective upcoming therapeutic approach might be directed to earlier disease stages.[10] Therefore, ideal surrogate markers should identify early disease-associated alterations before the hitherto unknown point of no return.[11]

In contrast to color fundus photography and fundus autofluorescence-based definition of GA,[12,13] the Classification of Atrophy Meetings (CAM) group (as an international consensus) recently used optical coherence tomography (OCT) imaging to redefine the phenotypic end stage of AMD as complete RPE and outer retinal atrophy (RORA). They not only included alterations of the outer retina into the definition but also reported preceding OCT features for AMD.[14,15] Furthermore, a current study dealing with RORA in mitochondriopathies described a consistent sequence of these OCT features in the development of RORA representing different disease stages.[16] Accordingly, they could bear great potential as future clinical surrogate markers. However, the reliability of the detection and quantification of some of these features has not yet been systematically and comprehensively investigated. Nevertheless, they have already been implemented by reading centers for current and upcoming observational and interventional trials.[14,17,18]

Concerning nAMD, the therapy with intraocular injection of anti–vascular endothelial growth factor (VEGF) has been shown to be effective and reduces the risk of visual loss.[19,20] However, the numbers and costs of required visits mean a significant burden on health care systems, medical personal, and patients, particularly in light of growing numbers due to demographic changes and rising life expectation.[21] Therefore, personalized interval and treatment strategies (i.e., "treat and extend") are used more commonly in current clinical settings.[22,23] In this context, objective and reliable features to determine disease activity are crucial. OCT is typically used for monitoring as it provides cross-sectional images of the retina that allow identifying the presence as well as extent of these features.[24,25] Usually, the feature identification is manually performed by human investigators. Machine learning (ML) applications are progressively entering this field, especially in the context of potential deployment of in-home or remote OCT monitoring.[26] However, the "gold standard" by which these algorithms are trained and validated is conventionally human grading. This might raise the question concerning reliability, subjectivity, and bias of the treatment decisions.[27]

In this study, we therefore investigate the reliability of the grading of defined OCT features commonly found in the development of RORA and/or in the presence of CNV secondary to AMD in order to provide estimates for human interreader agreement for each of these features. Thereby, we focus on the detection as well as the size and the overlap of the particular annotations.

Methods

This retrospective cross-sectional case series was performed at the Moorfields Eye Hospital NHS Foundation Trust (London, UK). To identify patients with AMD, the OCT images were linked to the diagnosis of the electronic medical records (EMR) database (Medisoft, Leeds, UK) of five centers in the United Kingdom using pseudonymized identifiers. The data pseudonymization was undertaken by the EMR vendor independently before export to the study team. The pseudonymization key that was generated to allow linkage of EMR to OCT data remained with the EMR vendor at the clinical site and not accessible to the study team, and all patient identifiers were removed. This means that the data received by the study team were effectively fully anonymized on receipt to prevent any possible identification of individual patients or treatment sites by the investigators. The imaging data comprised 6-mm × 6-mm foveal-centered OCT volume scans (128 or 256 scans per volume), resulting in a resolution of either 512 × 128 A-scans or 256 × 256 A-scans. They were obtained by spectral-domain OCT (Topcon, Tokyo, Japan) using standardized scan protocols. Any other additional ocular pathology (including prior clinically significant macular edema),

prior unlicensed bevacizumab injections, intraocular surgery within 90 days, or prior macular or panretinal photocoagulation led to exclusion. Thereby, this study included imaging data of 112 eyes of 112 patients with AMD at different disease stages. Half of these eyes were treatment naive, and the others were imaged after three anti-VEGF injections. Active neovascularization was present in 70 eyes. Of the remaining 42 eyes, 12 and 30 were graded as intermediate and late AMD, respectively. There were 60 right and 52 left eyes included. The mean ± SD age was 81.4 ± 8.18 years (range, 51–98 years). The study was in adherence with the Declaration of Helsinki. The institutional review board ruled that approval was not required for this study, because all data were effectively completely anonymized before being released to our study team to perform this research.

Image Analysis

To assess the reliability of grading retinal alterations in AMD, a single OCT B-scan per eye was randomly selected for annotation (including both foveal and eccentric scans). The other B-scans were available to give additional context if needed. Annotations were performed by four independently trained retinal specialists masked to the results of each other using a custom-build platform (Supplementary Fig. S1). All retinal abnormalities were to be delineated using (1) the definition of features as well as the images (as standard examples) of CAM reports[14,15] and (2) unpublished (additional) description of features based on the Classification of Atrophy Meeting from January 2019 in Milan, Italy (the corresponding CAM Report 5 is currently under review). The platform provided default labels for the most common abnormalities (including those described by the CAM group)[14,15] and allowed the readers to add additional labels not covered (as free text) by the default setup. The latter was used only once by one reader (annotating a single microaneurysm). Depending on the feature, it was annotated as area, lateral extent, or number (i.e., single dots in features with pointwise presentation) and likewise for all readers. Preset default labels included drusen, loss of ellipsoid zone (EZ), intraretinal hyperreflective foci (HRF), hypertransmission of OCT signal (HT), hyporeflective wedges, intraretinal fluid (IRF), descent of the outer plexiform layer (OPL), outer retinal tubulations, pigment epithelial detachment (PED), loss of retinal pigment epithelium (RPE), reticular pseudodrusen (RPD), subretinal fluid (SRF), subretinal hyperreflective material (SRHM), and sub-RPE plaques (Supplementary Fig. S1).

The annotated images were then evaluated using Python (version 3.8.2). To obtain the area measures in square millimeters and lateral extent measures in millimeters, the extracted values of annotated features (i.e., in pixels^2 and pixels) were multiplied by the individual scaling factor depending on the scanning protocol. Further statistical analysis was exclusively made for features present in at least 20 annotated B-scans (respectively, eyes) to ensure reliable results.

Statistical Analysis

The software environment R (version 4.0.2; The R Foundation for Statistical Computing, Vienna, Austria)[28] was used for interreader correlations. To compare the reliability of feature detection, Fleiss coefficients were used.[29] To measure the agreement in the annotated feature size, lateral extent, or number, intraclass correlation coefficients (ICCs, one-way random), 95% coefficients of repeatability, and coefficients of variation (CVs) were determined.[30–32] To account for the unbalanced number of readings per sample, a linear mixed-effects model was used. Bland–Altman plots were generated from slices with annotations of at least two readers for visualization of limits of agreement. Spearman's rank correlation coefficients (ρ) were calculated between the absolute differences and the mean values to evaluate whether measurement variability increases with lesion size or number.[31]

To measure overlap in annotated areas, we calculated the Dice similarity metric using Python (version 3.8.2; Python Software Foundation, Wilmington, Delaware, USA) whenever more than one reader annotated the same feature within a respective B-scan. It is defined as the size of the intersection of two areas divided by their average individual size, ranging from 0 (indicating no spatial overlap) to 1 (indicating complete overlap).[33] For area measures, overlap was calculated on the pixel level. For lateral extent measures, only the lateral location of the feature was taken into account. The mean Dice coefficients per feature are reported. Due to their focal nature, the Dice coefficient was not regarded an appropriate metric for annotations of HRF.

Results

In 111 of the included 112 OCT B-scans, at least one pathologic feature was annotated. Hyporeflective wedges ($n = 1$), microaneurysm ($n = 1$), outer retinal tubulations ($n = 5$), RPD ($n = 16$), and sub-RPE plaques ($n = 3$) were present but excluded from analy-

Table 1. Interreader Agreement of Feature Detection

Grading Parameter	*n*	κ Coefficient	95% CI
Drusen	85	0.367	0.292–0.443
Drusen_def.	64	0.613	0.537–0.689
EZ loss	108	0.260	0.185–0.336
HRF	71	0.422	0.246–0.497
HT	29	0.746	0.671–0.822
IRF	50	0.621	0.545–0.696
OPL descent	20	0.611	0.536–0.687
PED	77	0.598	0.522–0.674
RPE loss	76	0.160	0.085–0.236
SRF	45	0.823	0.747–0.898
SRHM	51	0.357	0.282–0.433

n = overall number of B-scans annotated with the respective feature by at least one reader. CI, confidence interval; Drusen_def., drusen with a minimum size of 1558.6 µm^2 in the respective B-scan.

Table 2. Interreader Agreement of Size, Lateral Extension, or Number of Annotated Features

Grading Parameter	CoR	CV, %	ICC (95% CI)
Drusen	0.098[a]	55.0	0.687 (0.534–0.792)
Drusen_def.	0.094[a]	48.5	0.788 (0.670–0.868)
EZ loss	3.446[b]	42.4	0.573 (0.415–0.695)
HRF_focality	9.388	64.5	0.527 (0.267–0.699)
HT	0.625[b]	24.1	0.936 (0.880–0.968)
IRF	0.121[a]	81.8	0.713 (0.525–0.831)
OPL descent	0.763[b]	16.2	0.884 (0.739–0.952)
PED	0.134[a]	17.6	0.972 (0.959–0.981)
RPE loss	2.157[b]	44.8	0.614 (0.345–0.766)
SRF	0.103[a]	46.5	0.938 (0.900–0.964)
SRHM	0.234[a]	53.9	0.793 (0.644–0.880)

CoR, 95% coefficients of repeatability; CV, Coefficients of variation; ICC; Intraclass correlation coefficients.

[a]Values indicate mm^2.

[b]Values indicate mm.

sis due to their rarity in the respective scans. In total, 10 features were used for further analysis (Table 1). Out of the latter group, EZ loss, drusen, and PED were the most abundant features.

The feature detection at the B-scan level (i.e., the individual lesion level is important when investigating progression) revealed variable interreader agreement (Table 1). The most reliable results could be found in SRF and IRF, which account for neovascular complications, as well as the features HT, OPL descent, and PED. Only slight to moderate interreader agreement could be found in the detection of EZ loss and RPE loss, quite similar to drusen grading.[29] However, setting a threshold of 1558.6 µm^2 as minimum drusen area (derived from the Age-Related Eye Disease Study (AREDS) definition of minimal drusen diameter of 63 µm)[34,35] to exclude so-called drupelets led to a reduced number of annotated B-scans ($n = 64$) and to a significantly increased κ coefficient of drusen grading, indicating substantial interreader agreement.

The evaluation of interreader agreement concerning the size, lateral extension, or number of annotated features at the B-scan level revealed more consistent results. All ICC values ranged from moderate to excellent correlation (Table 2).[36] The focality (i.e., number of individual annotated spots) measures of HRF revealed the lowest ICC with values over 0.50. The features with the highest scores for interreader agreement of annotated size, lateral extension, or number were PED, SRF, HT, and OPL descent in our cohort (ICC > 0.85, Fig. 1). Similar to the feature detection, exclusion of drupelets led to a higher interreader agreement of grading of drusen size (Table 2).

The Bland–Altman plots did not reveal systematic interreader discrepancies. Therefore, the mean difference between measurements by different readers consistently was around 0, and no pair of readers permanently showed higher or lower interreader agreement than the others (Fig. 2 and Supplementary Figs. S2–S11). However, the interreader variability increased with annotated area or number according to Spearman's rank correlation coefficient (ρ) for absolute differences and mean values for measures of drusen ($\rho = 0.317$ to $\rho = 0.828$, $P < 0.001$ to $P = 0.049$), PED ($\rho = 0.316$ to $\rho = 0.605$, $P < 0.001$ to $P = 0.042$), and HRF ($\rho = 0.509$ to $\rho = 0.761$, $P < 0.001$ to $P = 0.018$). The area measures of IRF ($\rho = 0.311$ to $\rho = 0.755$, $P < 0.001$ to $P = 0.139$), SRF ($\rho = 0.326$ to $\rho = 0.517$, $P = 0.003$ to $P = 0.062$), and SRHM ($\rho = 0.150$ to $\rho = 0.436$, $P = 0.170$ to $P = 0.708$), as well as lateral distance measures of EZ loss ($\rho = 0.010$ to $\rho = 0.297$, $P = 0.021$ to $P = 0.936$), HT ($\rho = 0.021$ to $\rho = 0.550$, $P = 0.027$ to $P = 0.921$), OPL descent ($\rho = 0.036$ to $\rho = 0.455$, $P = 0.066$ to $P = 0.964$), and RPE loss ($\rho = 0.108$ to $\rho = 0.748$, $P < 0.001$ to $P = 0.818$), did not show this correlation.

More reliable than size, extent, or number of annotated features, the Dice coefficients revealed consistent values over 0.5 (up to >0.75, Table 3) for all features. This indicated a distinct overlap of annotated regions and therefore uniform localization of the features (Fig. 1).

Discussion

In this study, we systematically investigated the reliability of grading an extensive number of structural OCT features associated with different stages of AMD

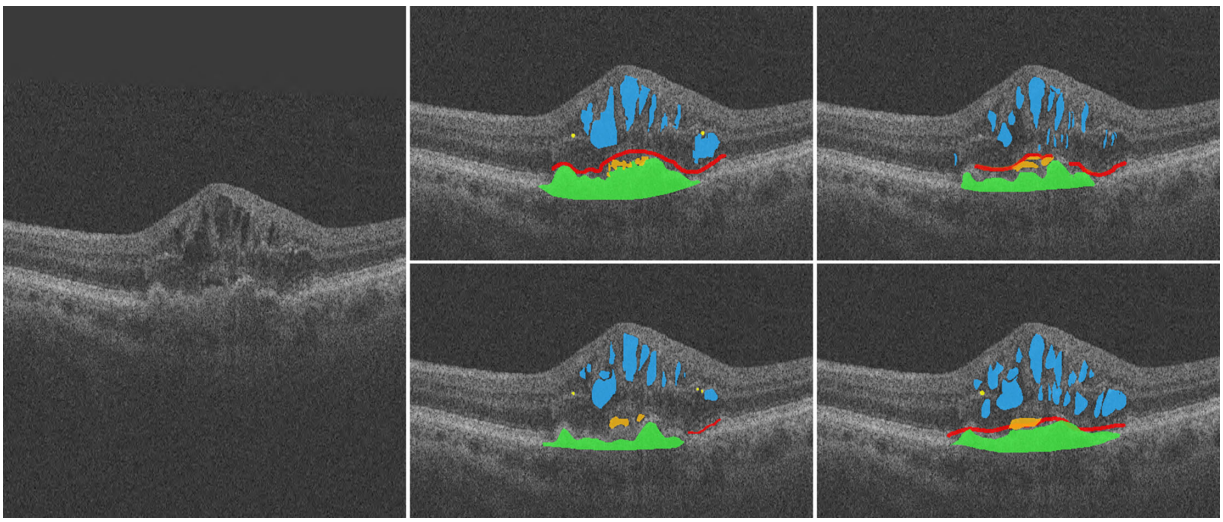

Figure 1. OCT-based feature annotation. An OCT B-scan (*left*) and the respective feature annotation of each reader (*right*) are demonstrated as example. IRF (*blue*), SRF (*orange*), and PED (*green*) revealed high interreader agreement, while annotations of EZ loss (*red*) and intraretinal HRF (*yellow*) significantly differed in size and number between the readers. However, the location of annotated features within the B-scan was quite similar throughout all features.

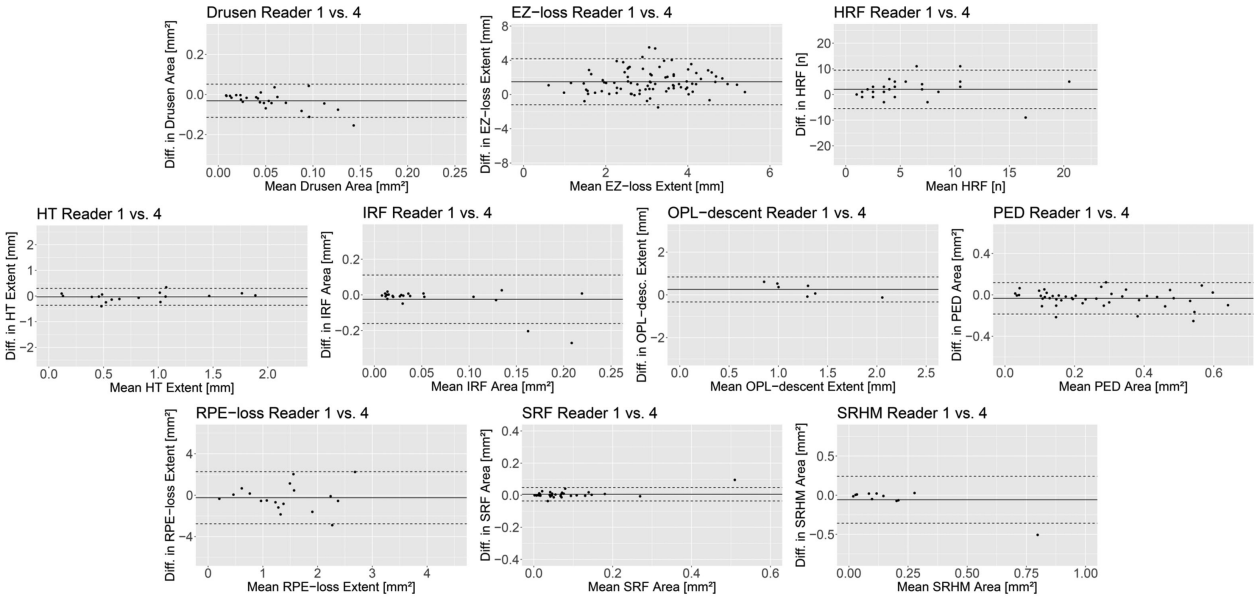

Figure 2. Interreader agreement. The Bland–Altman plots demonstrate the interreader agreement between two exemplary readers (readers 1 and 4) for measures of drusen, EZ loss, intraretinal HRF, HT, IRF, OPL descent, PED, RPE loss, SRF, and SRHM. The measurement differences (diff.) are plotted against their mean. The *solid line* indicates the mean difference and the *dashed lines* indicate the 95% limits of agreement. There were no systematic differences between the readers. Bland–Altman plots for the interreader agreement between each pair of all readers can be found in Supplementary Figures S2 to S11.

in a reading center setting. The presented findings provided evidence for the dependence of interreader agreement on the respective annotated feature. Hence, the appropriate selection of features has the potential to provide suitable surrogate markers for disease progression and possible therapeutic effects on different disease stages in upcoming interventional trials.

Clinical surrogate markers are needed to accelerate future interventional trials. Best-corrected visual acuity loss does not always constitute a useful endpoint in clinical trials for AMD due to its high interindividual variability, its psychophysical nature, and phenomena such as foveal noninvolvement.[37] Nevertheless, most interventional trials for neovascular AMD currently

Table 3. Interreader Agreement of Location of Annotated Features

Grading Parameter	Dice	95% CI
Drusen	0.539	0.507–0.570
EZ loss	0.632	0.606–0.658
HT	0.696	0.646–0.745
IRF	0.549	0.508–0.591
OPL descent	0.720	0.658–0.782
PED	0.764	0.740–0.787
RPE loss	0.650	0.598–0.701
SRF	0.664	0.632–0.697
SRHM	0.612	0.552–0.671

rely on this feature. In contrast, studies for dry AMD usually use morphologic endpoints like GA (e.g., by semiautomated delineation in fundus autofluorescence imaging)[38] or RORA (defined by OCT imaging)[14,15] as an accepted endpoint by regulators.[8,9] However, atrophic lesions represent the end stage of AMD, and the most effective upcoming therapeutic approach might be directed to earlier disease stages, which is difficult to extrapolate from preclinical data.[10] Ideal surrogate markers, therefore, should be readily captured, reflect the current disease stage, be reliable, and ideally be predictive for long-term progression based on short-term changes.[39]

As the OCT is the most abundant digital imaging device in modern ophthalmology, it has already been implemented in routine patient assessment and most clinical trial designs for retinopathies.[40] For neovascular AMD, the analysis of IRF and SRF is used to evaluate disease activity and treatment indication besides drop of vision, presence of bleedings, or leakage in angiography.[23,41] It has been shown to be an objective and susceptible measure that might even precede functional impairment and be faster executed and/or more comfortable than invasive imaging technology like angiography or fundus photography.[23,24,42] For dry AMD, multimodal assessment (including OCT) of drusen, pigment epithelial alterations, or signs of RORA is inevitable in the differential diagnosis and analysis of disease progression.[17] The evaluation of additional or individual OCT features could therefore be effectively carried out.

A current publication showed a consistent sequence of OCT features in the development of RORA secondary to maternally inherited diabetes and deafness (MIDD), indicating that these features represent different disease stages.[16] Given that MIDD is a mitochondriopathy and mitochondrial dysfunction is considered part of the pathophysiology in AMD,[43,44] results obtained in that model disease might be partly transferred to AMD. Indeed, an international consensus published by the CAM group indicated that most of these features are associated with RORA development secondary to AMD.[14,15] It also described features like EZ loss, RPE loss, HT, OPL descent, HRF, and SRHM. However, the reliability of these features has not yet been comprehensively investigated by this group.

Reliability might be the most important prerequisite to define a surrogate marker for patient assessment and future interventional clinical trials. Rather, low interreader agreement was found in the detection of EZ loss and RPE loss. Reliability of size and location of both feature annotations, however, were distinctly higher, while ICC did not reach levels of previously published data (0.75 for RPE loss).[45] However, the latter used another OCT device (Spectralis HRA-OCT; Heidelberg Engineering, Heidelberg, Germany) that might have led to better image quality. Some of the differences between readers might be due to inaccurate delineation of lesion borders since loss and attenuation of RPE and/or EZ might merge (Fig. 1). Interestingly, the average relative difference between two readers for RPE loss was indicated with 72.4, which was significantly higher than the CV (44.8) in our study, while both measures are thought to be independent of lesion size. Concerning HRF, the variable number might derive from the size of the feature. Readers might have simply overlooked small features, leading to not more than moderate reliability (Fig. 1). As these features with low interrater agreement might be inherently problematic for humans to detect and quantify on OCT images, their utility as surrogate markers in clinical studies is limited. In this context, an automated artificial intelligence–based feature detection is likely to be more consistent and precise in performance than human graders.[24,46,47] The application of deep learning and its broader family, ML, might be a way forward in utilizing the utility of these potential surrogate markers. However, the ML algorithms are trained and the performance is judged by the human "gold standard,"[48] which, if unreliable, may be problematic. Different approaches try to assess this problem: (1) Prerequisites for reliable gradings are precise definitions and grading protocols as well as proper annotation platforms (respectively, software environments). (2) Training a ML algorithm on gradings from multiple graders could converge these gradings to an average grader, which would mitigate part of the subjectivity.[49] (3) A consensus grading (e.g., from a consensus meeting or by averaging gradings or by adjudicating inconsistencies) might be considered "superhuman" (i.e., better than a single grader). This superhuman grading could be used to develop a model

that produces results at the same quality.[50] (4) The use of additional data (e.g., other modalities or follow-up images) may allow for improved grading.[51] (5) By using super-quality imaging (e.g., higher-resolution OCT), more reliable gradings might be obtained, which could then be transferred to standard-quality imaging for model development.[52] Moreover, ML is likely to be the only way to quantitate large volumes of dense OCT raster scans that are being generated in clinical trial reading centers, busy clinical practices, and emerging home/remote OCT devices.[53]

More consistent results could be found for SRF and IRF. Here, our results revealed high interreader agreement in all three investigated parameters (detection, size, and location; Fig. 1). This was in line with previously published data.[54] Despite different data sets, the here described ICCs between readers were higher than the ICCs derived from intermodality reliability between spectral-domain and time-domain OCT.[55,56] Given that both features reflect neovascular activity and guide the indication for anti-VEGF treatment (besides other clinical features, including hemorrhage and loss of vision), this might be of particular importance. A recent study has investigated the interreader agreement of PED size measures and reported an ICC of over 0.99.[57] The slightly higher ICC value (our study, 0.972) might be traced back to the fact that the latter has included only 20 eyes with a definite presence of PED and did not parallelly focus on other retinal alterations. The possible impact of reader fatigue (number of images and/or features) might be worth investigating in a future study.

We noted a high reliability of the HT feature, supporting previously published data.[45] In contrast, no previous report has systematically investigated interreader agreement of OPL descent. Given the high reliability (Tables 1–3) and appearance in the development of RORA,[14] OPL descent would be worth further investigations and to explore its potential as a possible surrogate marker in future clinical trials as well as for training ML algorithms.

Interestingly, the reliability of OCT-based feature annotation for SRF, HT, and PED, for example, reached the reliability of grading atrophic lesions in fundus autofluorescence imaging in different diseases, including AMD.[39,44,58,59] However, OCT imaging uses less energetic infrared light that minimizes potential light toxicity and is more comfortable for the patient.[58,60] Furthermore, OCT imaging does not rely on pupil dilation, and devices are more common than fundus autofluorescence imaging devices.[40] In this context, OCT scans were selected in a randomized manner in our study. A previous study revealed that more eccentric scan locations might lead to less reliable results.[54] Therefore, the pure evaluation of central scans might have led to even higher interreader agreement. Nevertheless, additional features of summation images like shape-descriptive parameters or dynamic flow signal could give further information,[44,59,61] suggesting a multimodal assessment as a gold standard in AMD diagnosis and study design at the current stage of imaging technology.[17]

It has been shown by the AREDS study that the number and size of drusen might predict progression of AMD.[62] Furthermore, we could show that the AREDS definition of minimum drusen size makes sense not only in the context of color fundus photography but also for OCT grading as the so-called drupelets (diameter <63 µm) have an unclear pathologic importance, and their exclusion led to a significant increase in interreader agreement (Tables 1 and 2).[34,35] If, nevertheless, a delineation of drupelets is aimed for, an automated artificial intelligence–based feature detection is likely to show improved performance over human graders, similar to the abovementioned small feature of HRF. More recently, a focus was set on the predictive value as well as the complicated delineation of drusen in the presence of RPD (also termed *subretinal drusenoid deposits*).[63] In this context, the low number of patients with RPD (which led to exclusion from further analysis) is a limitation of our study, and future studies focusing on interrater reliability of drusen, including this particular feature, are warranted. Besides drusen and RPD, the presentation of HRF[64] and the baseline atrophic lesion size[12] were also reported to affect future progression rate. Concerning exudative complications, the predictive value of SRF has been controversially discussed,[65,66] while the extent of central retinal thickening and IRF is thought to represent the neovascular activity and therefore visual outcome.[67–69] Therefore, it might be hypothesized that some of the additionally presented imaging features could also be predictive for neovascular or dry AMD progression. However, the image feature description in this study was based on retrospective cross-sectional data, as it was beyond the scope of this study to evaluate the accuracy of predictive factors. However, if noted to be present, the consistency of size and location of most imaging features have the potential to provide the framework for further prospective studies. These prospective studies would allow to further evaluate the predictive value, which might give more insights into the pathophysiology of AMD and allow for effective study design as presented before for different parameters in AMD or other retinopathies.[42,44,59,61,70,71]

A further limitation of this study is the application of OCT imaging devices by a single manufacturer. Different OCT imaging devices might provide different

scanning artifacts or image quality.[72,73] Thereby, the annotation and, hence, the reliability of single features might be different on large-scale real-world data.[74] As there is no gold standard, it cannot be excluded that features have been missed and other data sets could provide addition conclusions. To minimize this possibility, we relied on trained retinal specialists who have identified and interpreted the features, and the opportunity to add additional features was given at all time points during annotation (Supplementary Fig. S1). Finally, readers might have utilized the contextual B-scans differently, which was not recorded. However, the variability of their approach and annotations reflects the human variability, which was part of the purpose of this study. An evaluation of how human readers use additional images for grading might be an interesting question for a future study, especially in the context of multimodal approaches to retinal diseases.[17]

In conclusion, this study evaluated the reliability of annotations of multiple OCT features representing different disease stages in a reading center setup. The inclusion of objective and reliable features like SRF, IRF, HT, OPL descent, or PED into future studies might enable multiple surrogate markers representing different disease stages within a single image. This might open up numerous new opportunities for evaluating disease progression and possible treatment effect in AMD, possibly leading to a more time- and cost-effective interpretation, further insights into the pathomechanisms, enhanced individualized patient assessment, and improved training of ML application. Emerging advances in artificial intelligence training and validation may allow for a higher consistency in performance than human graders, suggesting a wider variety of reliable surrogate markers and potential benefits in the future.

Acknowledgments

The authors thank the members of the CAM group for setting the standards of the feature grading for this work. This work was supported by the German Research Foundation (grant MU4279/2-1 to PLM), the United Kingdom's National Institute for Health Research of Health's Biomedical Research Centre for Ophthalmology at Moorfields Eye Hospital, and UCL Institute of Ophthalmology. The views expressed are those of the authors and not necessarily those of the Department of Health. The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Disclosure: **P.L. Müller,** None; **B. Liefers,** None; **T. Treis,** None; **F.G. Rodrigues,** None; **A. Olvera-Barrios,** None; **B. Paul,** None; **N. Dhingra,** None; **A. Lotery,** None; **C. Bailey,** None; **P. Taylor,** None; **C.I. Sánchez,** None; **A. Tufail,** None

References

1. Flaxman SR, Bourne RRA, Resnikoff S, et al. Global causes of blindness and distance vision impairment 1990–2020: a systematic review and meta-analysis. *Lancet Glob Heal*. 2017;5(12):e1221–e1234.
2. Holz FG, Pauleikhoff D, Klein R, Bird AC. Pathogenesis of lesions in late age-related macular disease. *Am J Ophthalmol*. 2004;137(3):504–510.
3. Khan M, Agarwal K, Loutfi M, Kamal A. Present and possible therapies for age-related macular degeneration. *ISRN Ophthalmol*. 2014;2014:1–7.
4. Holz FG, Schmitz-Valckenberg S, Fleckenstein M. Recent developments in the treatment of age-related macular degeneration. *J Clin Invest*. 2014;124(4):1430–1438.
5. Holz FG, Sadda SR, Busbee B, et al. Efficacy and safety of lampalizumab for geographic atrophy due to age-related macular degeneration. *JAMA Ophthalmol*. 2018;136(6):666–677.
6. Rosenfeld PJ, Dugel PU, Holz FG, et al. Emixustat hydrochloride for geographic atrophy secondary to age-related macular degeneration. *Ophthalmology*. 2018;125(10):1556–1567.
7. Terheyden JH, Holz FG, Schmitz-Valckenberg S, et al. Clinical study protocol for a low-interventional study in intermediate age-related macular degeneration developing novel clinical endpoints for interventional clinical trials with a regulatory and patient access intention—MACUSTAR. *Trials*. 2020;21(1):659.
8. Csaky KG, Richman EA, Ferris FL. Report from the NEI/FDA Ophthalmic Clinical Trial Design and Endpoints Symposium. *Invest Ophthalmol Vis Sci*. 2008;49(2):479–489.
9. Holz FG, Strauss EC, Schmitz-Valckenberg S, van Lookeren Campagne M. Geographic atrophy: clinical features and potential therapeutic approaches. *Ophthalmology*. 2014;121(5):1079–1091.
10. Schaal KB, Rosenfeld PJ, Gregori G, Yehoshua Z, Feuer WJ. Anatomic clinical trial endpoints

for nonexudative age-related macular degeneration. *Ophthalmology*. 2016;123(5):1060–1079.
11. Finger RP, Schmitz-Valckenberg S, Schmid M, et al. MACUSTAR: development and clinical validation of functional, structural, and patient-reported endpoints in intermediate age-related macular degeneration. *Ophthalmologica*. 2019;241(2):61–72.
12. Lindblad AS, Lloyd PC, Clemons TE, et al. Change in area of geographic atrophy in the age-related eye disease study: AREDS report number 26. *Arch Ophthalmol*. 2009;127(9):1168–1174.
13. Holz FG, Strauss EC, Schmitz-Valckenberg S, van Lookeren Campagne M. Geographic atrophy. *Ophthalmology*. 2014;121(5):1079–1091.
14. Sadda SR, Guymer R, Holz FG, et al. Consensus definition for atrophy associated with age-related macular degeneration on OCT: Classification of Atrophy Report 3. *Ophthalmology*. 2018;125(4):537–548.
15. Guymer RH, Rosenfeld PJ, Curcio CA, et al. Incomplete retinal pigment epithelial and outer retinal atrophy in age-related macular degeneration: Classification of Atrophy Meeting Report 4. *Ophthalmology*. 2020;127(3):394–409.
16. Müller PL, Maloca P, Webster A, Egan C, Tufail A. Structural features associated with the development and progression of RORA secondary to maternally inherited diabetes and deafness. *Am J Ophthalmol*. 2020;218:136–147.
17. Holz FG, Sadda SR, Staurenghi G, et al. Imaging protocols in clinical studies in advanced age-related macular degeneration: recommendations from classification of atrophy consensus meetings. *Ophthalmology*. 2017;124(4):464–478.
18. Guymer RH, Wu Z, Hodgson LAB, et al. Subthreshold nanosecond laser intervention in age-related macular degeneration: the LEAD randomized controlled clinical trial. *Ophthalmology*. 2019;126(6):829–838.
19. Rofagha S, Bhisitkul RB, Boyer DS, Sadda SR, Zhang K, Study Group SEVEN-UP. Seven-year outcomes in ranibizumab-treated patients in ANCHOR, MARINA, and HORIZON. *Ophthalmology*. 2013;120(11):2292–2299.
20. María GM, Paz SR, Isabel FRM, et al. Pharmacological advances in the treatment of age-related macular degeneration. *Curr Med Chem*. 2019;26:1–5.
21. Wong WL, Su X, Li X, et al. Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. *Lancet Glob Heal*. 2014;2(2):e106–e116.
22. Holz FG, Amoaku W, Donate J, et al. Safety and efficacy of a flexible dosing regimen of ranibizumab in neovascular age-related macular degeneration: the SUSTAIN study. *Ophthalmology*. 2011;118(4):663–671.
23. Lee A, G Garg P, T Lyon A, Mirza R, K Gill M. Long-term outcomes of treat and extend regimen of anti–vascular endothelial growth factor in neovascular age-related macular degeneration. *J Ophthalmic Vis Res*. 2020;15(3):331–340.
24. Schmidt-Erfurth U, Klimscha S, Waldstein SM, Bogunović H. A view of the current and future role of optical coherence tomography in the management of age-related macular degeneration. *Eye*. 2017;31(1):26–44.
25. Waldstein SM, Philip A-M, Leitner R, et al. Correlation of 3-dimensionally quantified intraretinal and subretinal fluid with visual acuity in neovascular age-related macular degeneration. *JAMA Ophthalmol*. 2016;134(2):182.
26. Quellec G, Kowal J, Hasler PW, et al. Feasibility of support vector machine learning in age-related macular degeneration using small sample yielding sparse optical coherence tomography data. *Acta Ophthalmol*. 2019;97(5):e719–e728.
27. Toth CA, Decroos FC, Ying G-S, 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(7):1303–1314.
28. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2015.
29. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159–174.
30. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420–428.
31. Bland JM, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;327(8476):307–310.
32. Mair G, von Kummer R, Adami A, et al. Observer reliability of CT angiography in the assessment of acute ischaemic stroke: data from the Third International Stroke Trial. *Neuroradiology*. 2015;57(1):1–9.
33. Dice LR. Measures of the amount of ecologic association between species. *Ecology*. 1945;26(3):297–302.
34. Age-Related Eye Disease Study Research Group. The age-related eye disease study system for

classifying age-related macular degeneration from stereoscopic color fundus photographs: the Age-Related Eye Disease Study report number 6. *Am J Ophthalmol*. 2001;132(5):668–681.
35. Ferris FL, Wilkinson CP, Bird A, et al. Clinical classification of age-related macular degeneration. *Ophthalmology*. 2013;120(4):844–851.
36. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2):155–163.
37. Schmitz-Valckenberg S, Fleckenstein M, Helb H-M, Issa PC, Scholl HPN, Holz FG. In vivo imaging of foveal sparing in geographic atrophy secondary to age-related macular degeneration. *Invest Ophthalmol Vis Sci*. 2009;50(8):3915.
38. Schmitz-Valckenberg S, Brinkmann CK, Alten F, et al. Semiautomated image processing method for identification and quantification of geographic atrophy in age-related macular degeneration. *Invest Ophthalmol Vis Sci*. 2011;52(10):7640–7646.
39. Pfau M, Goerdt L, Schmitz-Valckenberg S, et al. Green-light autofluorescence versus combined blue-light autofluorescence and near-infrared reflectance imaging in geographic atrophy secondary to age-related macular degeneration. *Invest Ophthalmol Vis Sci*. 2017;58(6):121–130.
40. Müller PL, Wolf S, Dolz-Marco R, Tafreshi A, Schmitz-Valckenberg S, Holz FG. Ophthalmic diagnostic imaging: retina. In: Bille JF, ed. *High Resolution Imaging in Microscopy and Ophthalmology: New Frontiers in Biomedical Optics*. Springer International Publishing, Cham (CH); 2019:87–106.
41. Rosenfeld PJ. Optical coherence tomography and the development of antiangiogenic therapies in neovascular age-related macular degeneration. *Invest Ophthalmol Vis Sci*. 2016;57(9):OCT14.
42. Sleiman K, Veerappan M, Winter KP, et al. Optical coherence tomography predictors of risk for progression to non-neovascular atrophic age-related macular degeneration. *Ophthalmology*. 2017;124(12):1764–1777.
43. Suter M, Remé C, Grimm C, et al. Age-related macular degeneration: the lipofusion component N-retinyl-N-retinylidene ethanolamine detaches proapoptotic proteins from mitochondria and induces apoptosis in mammalian retinal pigment epithelial cells. *J Biol Chem*. 2000;275(50):39625–39630.
44. Müller PL, Treis T, Pfau M, et al. Progression of retinopathy secondary to maternally inherited diabetes and deafness—evaluation of predicting parameters. *Am J Ophthalmol*. 2020;213:134–144.
45. Sayegh RG, Simader C, Scheschy U, et al. A systematic comparison of spectral-domain optical coherence tomography and fundus autofluorescence in patients with geographic atrophy. *Ophthalmology*. 2011;118(9):1844–1851.
46. Maloca PM, Lee AY, de Carvalho ER, et al. Validation of automated artificial intelligence segmentation of optical coherence tomography images. *PLoS One*. 2019;14(8):e0220063.
47. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Heal*. 2019;1(6):e271–e297.
48. Pfau M, Walther G, von der Emde L, et al. Artificial intelligence in ophthalmology: guidelines for physicians for the critical evaluation of studies. *Ophthalmologe*. 2020;117(10):973–988.
49. Liefers B, Colijn JM, González-Gonzalo C, et al. A deep learning model for segmentation of geographic atrophy to study its long-term natural history. *Ophthalmology*. 2020;127(8):1086–1096.
50. Krause J, Gulshan V, Rahimy E, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*. 2018;125(8):1264–1272.
51. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med*. 2018;24(9):1342–1350.
52. Venhuizen FG, van Ginneken B, Liefers B, et al. Deep learning approach for the detection and quantification of intraretinal cystoid fluid in multivendor optical coherence tomography. *Biomed Opt Express*. 2018;9(4):1545.
53. Lee A, Taylor P, Kalpathy-Cramer J, Tufail A. Machine learning has arrived! *Ophthalmology*. 2017;124(12):1726–1728.
54. Kashani AH, Keane PA, Dustin L, Walsh AC, Sadda SR. Quantitative subanalysis of cystoid spaces and outer nuclear layer using optical coherence tomography in age-related macular degeneration. *Investig Opthalmology Vis Sci*. 2009;50(7):3366.
55. Folgar FA, Jaffe GJ, Ying G-S, Maguire MG, Toth CA; Comparison of Age-Related Macular Degeneration Treatments Trials Research Group. Comparison of optical coherence tomography assessments in the comparison of age-related macular degeneration treatments trials. *Ophthalmology*. 2014;121(10):1956–1965.

56. Cukras C, Wang YD, Meyerle CB, Forooghian F, Chew EY, Wong WT. Optical coherence tomography–based decision making in exudative age-related macular degeneration: comparison of time- vs spectral-domain devices. *Eye (Lond)*. 2010;24(5):775–783.
57. Ohayon A, Semoun O, Caillaux V, et al. Reliability and reproducibility of pigment epithelial detachment volume measurements in AMD using a new tool: ReVAnalyzer. *Ophthalmic Surg Lasers Imaging Retina*. 2019;50(9):e242–e249.
58. Müller PL, Pfau M, Mauschitz MM, et al. Comparison of green versus blue fundus autofluorescence in ABCA4-related retinopathy. *Transl Vis Sci Technol*. 2018;7(5):13.
59. Müller PL, Pfau M, Treis T, et al. Progression of ABCA4-related retinopathy—prognostic value of demographic, functional, genetic and imaging parameters [published online January 8, 2020]. *Retina*.
60. Müller PL, Birtel J, Herrmann P, Holz FG, Charbel Issa P, Gliem M. Functional relevance and structural correlates of near infrared and short wavelength fundus autofluorescence imaging in ABCA4-related retinopathy. *Transl Vis Sci Technol*. 2019;8(6):46.
61. Pfau M, Lindner M, Goerdt L, et al. Prognostic value of shape-descriptive factors for the progression of geographic atrophy secondary to age-related macular degeneration. *Retina*. 2019;39(8):1527–1540.
62. Chew EY, Clemons TE, Agrón E, et al. Ten-year follow-up of age-related macular degeneration in the age-related eye disease study: AREDS report no. 36. *JAMA Ophthalmol*. 2014;132(3):272–277.
63. Sadda SR, Abdelfattah NS, Lei J, et al. Spectral-domain OCT analysis of risk factors for macular atrophy development in the HARBOR study for neovascular age-related macular degeneration. *Ophthalmology*. 2020;127(10):1360–1370.
64. Schmidt-Erfurth U, Bogunovic H, Grechenig C, et al. Role of deep learning quantified hyperreflective foci for the prediction of geographic atrophy progression. *Am J Ophthalmol*. 2020;216:257–270.
65. Schmidt-Erfurth U, Waldstein SM. A paradigm shift in imaging biomarkers in neovascular age-related macular degeneration. *Prog Retin Eye Res*. 2016;50:1–24.
66. Arnold JJ, Markey CM, Kurstjens NP, Guymer RH. The role of sub-retinal fluid in determining treatment outcomes in patients with neovascular age-related macular degeneration—a phase IV randomised clinical trial with ranibizumab: the FLUID study. *BMC Ophthalmol*. 2016;16(1):31.
67. Rosenfeld PJ, Brown DM, Heier JS, et al. Ranibizumab for neovascular age-related macular degeneration. *N Engl J Med*. 2006;355(14):1419–1431.
68. Simader C, Ritter M, Bolz M, et al. Morphologic parameters relevant for visual outcome during anti-angiogenic therapy of neovascular age-related macular degeneration. *Ophthalmology*. 2014;121(6):1237–1245.
69. Ying G, Huang J, Maguire MG, et al. Baseline predictors for one-year visual outcomes with ranibizumab or bevacizumab for neovascular age-related macular degeneration. *Ophthalmology*. 2013;120(1):122–129.
70. Thiele S, Nadal J, Pfau M, et al. Prognostic value of retinal layers in comparison with other risk factors for conversion of intermediate age-related macular degeneration. *Ophthalmol Retin*. 2020;4(1):31–40.
71. Müller PL, Treis T, Odainic A, et al. Prediction of function in ABCA4-related retinopathy using ensemble machine learning. *J Clin Med*. 2020;9(8):2428.
72. Tan CS, Chan JC, Cheong KX, Ngo WK, Sadda SR. Comparison of retinal thicknesses measured using swept-source and spectral-domain optical coherence tomography devices. *Ophthalmic Surg Lasers Imaging Retina*. 2015;46(2):172–179.
73. Mitsch C, Lammer J, Karst S, Scholda C, Pablik E, Schmidt-Erfurth UM. Systematic ultrastructural comparison of swept-source and full-depth spectral domain optical coherence tomography imaging of diabetic macular oedema. *Br J Ophthalmol*. 2020;104(6):868–873.
74. Al-Sheikh M, Ghasemi Falavarjani K, Akil H, Sadda SR. Impact of image quality on OCT angiography based quantitative measurements. *Int J Retin Vitr*. 2017;3:13.