Article

Reproducibility of Measurements of Retinal Structural Parameters Using Optical Coherence Tomography in Stargardt Disease

Xiangrong Kong[1,2], Alexander Ho[3], Beatriz Munoz[1], Sheila West[1], Rupert W. Strauss[1,4,5,6], Anamika Jha[3], Ann Ervin[1,2], Jeff Buzas[7], Mandeep Singh[1], Zhihong Hu[3], Janet Cheetham[8], Michael Ip[3], and Hendrik P. N. Scholl[1,9,10]

[1] Wilmer Eye Institute, Johns Hopkins University, Baltimore, MD, USA
[2] Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA
[3] Doheny Eye Institute, David Geffen School of Medicine at University of California Los Angeles, CA, USA
[4] Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, University College London, London, UK
[5] Department of Ophthalmology, Johannes Kepler University Clinic Linz, Linz, Austria
[6] Department of Ophthalmology, Medical University Graz, Graz, Austria
[7] Department of Statistics, University of Vermont, Burlington, VM, USA
[8] Foundation Fighting Blindness, Columbia, MD, USA
[9] Institute of Molecular and Clinical Ophthalmology Basel (IOB), Basel, Switzerland
[10] Department of Ophthalmology, University of Basel, Basel, Switzerland

Correspondence: Hendrik P. N. Scholl, Institute of Molecular and Clinical Ophthalmology Basel (IOB), Mittlere Strasse 91, CH-4031 Basel, Switzerland. e-mail: Hendrik.Scholl@iob.ch

Purpose: To assess the reproducibility of retinal measurements from optical coherence tomography (OCT) in *ABCA4*-related Stargardt disease (STGD1).

Methods: The international multicenter Progression of Atrophy Secondary to Stargardt Disease (ProgStar) Study enrolled 259 STGD1 patients. OCT images were graded by the study reading center (RC). Semiautomatic segmentation with manual adjustments was used to segment the layers of retinal pigmentation epithelium, outer segments, inner segments (ISs), outer nuclear layer (ONL), inner retina, and the total retina (TR). The images were overlaid to the Early Treatment Diabetic Retinopathy Study (ETDRS) grid. For each layer, the thickness and the intact area of the ETDRS central subfield, inner ring, and outer ring were recorded, respectively. A different set of RC graders regraded 30 independent ProgStar images to evaluate measurement reproducibility. Reproducibility was assessed graphically and using statistics including intraclass correlation (ICC) and relative absolute difference (RAD).

Results: Across all layers, measurements of the ETDRS central subfield had low ICC and/or large RAD. The outer-ring region was not fully captured in some images. For inner ring, good reproducibility was observed for intact area in the IS (ICC = 0.99, RAD = 4%), thicknesses of the ONL (ICC = 0.93, RAD = 6%), and TR (ICC = 0.99, RAD = 1%).

Conclusions: STGD1's complex morphology made outer retina segmentation challenging. Measurements of the inner ring, including the intact area of IS (i.e., the ellipsoid zone [EZ]) and ONL and TR thicknesses, had good reproducibility and showed anatomical impairment.

Translational Relevance: ONL and TR thicknesses and the EZ intact area in the ETDRS inner ring hold potential as structural endpoints for STGD1 trials. Structure-function relationships need to be further established.

Introduction

Stargardt disease is the most common macular dystrophy affecting both children and adults[1] and is inherited as an autosomal-recessive trait associated with mutations in the *ABCA4* gene (STGD1; OMIM: 248200). Clinically, it is characterized by fundus flecks in the retinal pigment epithelium (RPE) and by macular atrophic lesions. Patients experience slow progressive loss of retinal function, especially a loss of central vision. The loss of RPE and photoreceptor cells and the associated retinal thinning is hypothesized to be the morphological change in STGD1 leading to loss of visual function.

Currently there is no treatment for STGD1.[1–7] The international multicenter Progression of Atrophy Secondary to Stargardt Disease (ProgStar) study aims to understand the natural history of disease progression and determine appropriate outcome measures for future treatment trials of STGD1.[6] Spectral-domain optical coherence tomography (SD-OCT) represents a major imaging modality in the ProgStar study. Using SD-OCT to measure different retinal structural parameters, especially the thickness and intactness of RPE and the photoreceptor outer segment (OS) and inner segment (IS) in the fovea and parafovea regions, may offer an objective way to document the structural changes of STGD1.

SD-OCT is a noninvasive imaging technology and is becoming increasingly indispensable in ophthalmologic research and clinical practice.[8] Through assessing high-resolution, cross-sectional images of the posterior segment, SD-OCT generates quantitative measures of structural parameters of the posterior segment and, thus, provides a powerful tool for monitoring changes in ocular structure and for quantifying the rate of disease progression. However, to draw valid inference by using the quantitative measurements from SD-OCT, it is necessary that the measurements are reliable, reflecting the true state of the ocular structure. Nevertheless, in practice, a myriad of factors can impact the reliability of SD-OCT measurements, such as image quality, photographer and grader experience, patient's cooperation, and potential confounding pathology. Therefore, obtaining accurate and precise measurements (i.e., unbiased measurements with low variability) from SD-OCT is challenging, and it is crucial to evaluate the reliability of a SD-OCT measurement before using it to track progression of pathology.

To help determine whether retinal structural measurements from SD-OCT can provide unbiased estimates of rates of disease progression in STGD1 and whether the measurements may serve as endpoints for future trials, we report the reproducibility[9] assessment (i.e., repeated grading of the same images by a different set of graders) of the SD-OCT measurements from the ProgStar study and discuss the challenges in SD-OCT grading for STGD1.

Methods

Participants

The prospective ProgStar study (www.clinicaltrials. gov registration NCT01977846) was approved by the Western Institutional Review Board (IRB), local IRBs, and the Human Research Protection Office of the United States Army Medical Research and Materiel Command. The Declaration of Helsinki was adhered to throughout the study. Details of the prospective ProgStar study have been described elsewhere.[6] In brief, 259 participants were enrolled at nine sites from September 2013 to March 2015. Eligibility criteria included the following: age of ≥ 6 years and having two pathogenic mutations in the *ABCA4* gene or having one pathogenic mutation in the *ABCA4* gene at the time of genetic testing plus a typical Stargardt phenotype.[6] Study eyes further required that there was at least one well-demarcated area of atrophy with a minimum diameter of 300 μm and that the total area of atrophy was 12 mm^2 or less, as assessed by the site investigator using fundus autofluorescence imaging. The study eyes also needed to have best corrected visual acuity (BCVA) of 20 ETDRS letters or better at 1 m (i.e. Snellen equivalent 20/400 or better). All participants gave written informed consent prior to enrollment and were followed semiannually for 2 years. SD-OCT measurements were obtained for images captured at the baseline and 6-, 12-, and 24-month visits, and the grading occurred during 2014–2018.

Reproducibility of SD-OCT measurements was assessed by regrading 30 independent images from the prospective ProgStar study. The 30 images were randomly selected from images captured at the baseline or 6-, 12-, or 24-month visits and were from 30 different participants. Specifically, the 30 images were drawn from images that had all relevant retinal layers present in the inner ring area of the ETDRS grid centered at the fovea (Fig. 1). The regrading occurred in January 2018, and different sets of graders performed the original grading and regrading.
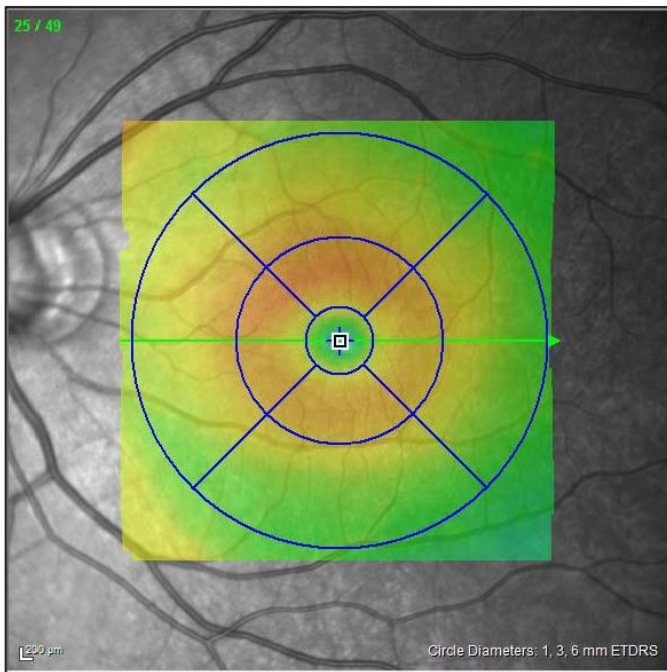
Figure 1. ETDRS grid: the central subfield is the circle in the center with a diameter of 1 mm. The inner ring consists of the four subfields surrounding the central circle, and the middle circle has a diameter of 3 mm. The outer ring consists of the four subfields surrounding the inner ring, and the outer circle has a diameter of 6 mm.

The regrading was masked from results obtained from the original grading.

SD-OCT Image Capturing and Grading

The Doheny Image Reading Center (DIRC) was the ProgStar reading center. SD-OCT images were captured using DIRC-approved Heidelberg Spectralis SD-OCT instruments and by DIRC-qualified photographers. A 20° × 20°, 49-section high-resolution infrared (IR)+OCT volume scan (enhanced depth imaging [EDI]) off centered on the fovea was acquired from 30° field of view and was uploaded directly to DIRC's data server. Semiautomatic segmentation of retinal layers started with an algorithm implemented in a proprietary grading software 3D-OCTOR.[10] Graders then manually adjusted the segmentation to correct for errors. The segmented boundaries included inner limiting membrane, inner aspect of the outer plexiform layer, external limiting membrane (ELM), IS/OS junction (also called the ellipsoid zone [EZ]), outer photoreceptor segment layer, inner RPE cell layer, and inner choroid (Fig. 2). If graders determined that a B-scan added little information to the segmentation compared to interpolating between neighboring segmented B-scans, the graders may

choose not to segment this B-scan, provided that at least 24 scans out of the total 49 B-scans must be graded per eye and no more than 4 consecutive scans can be excluded from segmentation. Following segmentation, graders also manually centered the ETDRS grid on the anatomical foveal center. On average, grading of one image required 6 hours of manual segmentation.

Based on the segmentation and overlaying with the ETDRS grid, two outcomes, the thickness and intact area, were measured for each of six retinal layers in the central subfield, inner ring, outer ring, and the total ETDRS grid, respectively (Figs. 1, 2). The six retinal layers are the following: RPE, between the inner RPE cell layer and the inner choroid; OS, between the IS/OS junction (EZ) and the outer photoreceptor segment layer; IS, the ELM and the IS/OS junction (EZ); outer nuclear layer (ONL), between the inner aspect of the outer plexiform layer and the ELM (the inner aspect of the outer plexiform layer was used to avoid the challenge in distinguishing the Henle fiber layer from the ONL with off-axis OCT scans[11]); inner retina, between the inner limiting membrane and the inner aspect of the outer plexiform layer; and the total retina, between the inner limiting membrane and the outer photoreceptor segment layer. If the EZ was disrupted, the boundaries above and below (the ELM, the outer photoreceptor segment layer) and the IS/OS junction were artificially snapped together (Supplementary Fig. S1). In such case, the thickness and intact area for the IS and OS layers would be all 0.

Two graders, including at least one senior grader, graded each image. Due to the significant effort needed for the manual segmentation process, grading from the two graders was not masked from each other: the first grader must consult the second grader on segmentation strategy (e.g., presence of layers; segmentation versus interpolation strategy). Segmentation was then performed by the first grader and then reviewed by the second grader. In the event that the two graders were not able to agree on a final answer after adjudication, a reading center investigator determined the final result.

Statistical Analysis

The demographic and clinical characteristics of the 30 participants (eyes) were summarized. Reproducibility of the OCT measurements was evaluated separately for the two outcome measures (thickness and intact area) for each of the six aforementioned retinal layers and for the ETDRS central subfield, inner ring and outer ring, respectively. Reproducibil-
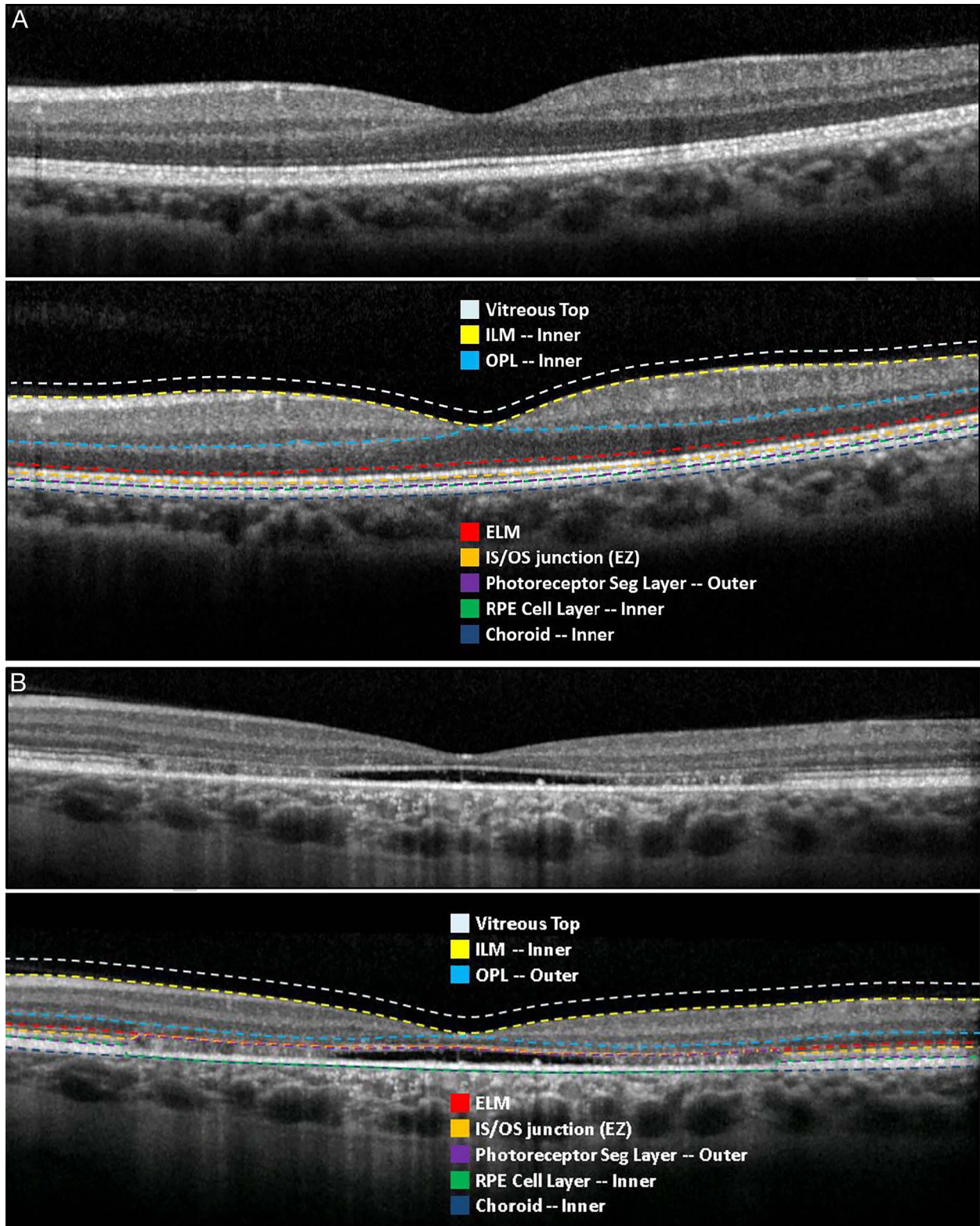
Figure 2. Retinal boundaries and layers in SD-OCT segmentation adopted in the ProgStar study. (A) An example of segmentation in a normal eye. Note, the photoreceptor segment layer, namely, outer (*purple*) is snapped to the RPE cell layer, and inner (*green*) boundary to denote no thickness due to subretinal fluid/subretinal hyperreflective materials. (B) An example of segmentation in a Stargardt patient eye. Note relevant boundaries are snapped together at locations where IS/OS junction is not present, subretinal

→

←

fluid/subretinal hyperreflective materials are present, and/or RPE is atrophied. ILM, inner limiting membrane; OPL, outer plexiform layer; ELM, external limiting membrane; IS/OS junction, inner segment-outer segment junction, called ellipsoid zone (EZ); RPE, retinal pigment epithelium.

ity for each parameter was assessed in four ways: first, graphically by the Bland-Altman (B-A) plots including estimating the 95% limits of agreement[12]; second, by pairwise comparisons between the two gradings using paired *t*-tests; third, by estimating the relative absolute difference between gradings (i.e., the absolute value of [grading 1 − grading 2] divided by the average of the two gradings); and fourth, by estimating the intraclass correlations (ICCs) using the Shrout and Fleiss formula, which is used when each image is rated by multiple and the same number of raters, assuming rater effect is random.[13]

Additionally, for the outcome measure of thickness, the prevalence of loss of a layer (i.e., thickness of the layer = 0) was estimated for each layer for the ETDRS central subfield, and the percentage of agreement and Kappa statistics were used to assess the agreement between the two gradings.[14] Because when thickness was measured as 0, the intact area was also 0, agreement assessment was not duplicated for the measure of intact area.

All analysis was conducted in SAS 9.3 (Cary, NC) and used the intraclass SAS Macro.[15] Two-sided *P* values were reported when *P* values were relevant.

Results

Demographics of the 30 participants and the clinical characteristics of the eyes selected for the reproducibility study are summarized in Table 1 (the corresponding demographic and clinical profiles of the whole ProgStar cohort is also presented in Tables 1 and 2). The median age was 32 years, 47% (*n* = 14) female, and 13% (*n* = 4) were non-whites. The median BCVA was 0.78 (range, 0.08–1.16) LogMAR. Table 3 shows summaries of the original grading measurements of the thickness and intact area of each retinal layer by ETDRS region. For example, the mean thickness of the total retina was 148.1 μm in the ETDRS central subfield and 248.1 μm and 274.4 μm in the inner and outer ring, respectively. The mean intact area of the total retina in the ETDRS center and inner and outer ring was 0.79, 6.28, and 20.78 mm^2, respectively. Table 3 also shows that in the central subfield, the prevalence of the loss of RPE, OS, and IS was 33%, 70%, and 73%, respectively.

Retinal Layer: RPE

Figure 3A presents the B-A plots of the thickness and intact area of the RPE in the ETDRS central subfield, inner ring, and outer ring. In each plot, the x axis (average of the two gradings) shows the magnitude of the thickness/intact area of the layer, and the y axis reflects the level of difference between the two gradings on the variable. For the measure of RPE thickness, although there is no significant mean difference between gradings in each ETDRS region, there is large variability on the difference between gradings, as evidenced by the wide range between the limits of agreement in each plot, the poor ICCs (0.76, 0.73, and 0.63 in center, inner ring, and outer ring, respectively; Table 4), and also by the high relative absolute differences (RADs) between gradings (14% for the inner ring and 11% for the outer ring; Table 4). The B-A plots also showed potential heteroscedastic difference between gradings where the difference between grading depended on the magnitude of thickness itself. Additionally, assessing agreement on thickness as a dichotomized outcome (>0 vs. =0), the percentage of agreement was 70%, and the Kappa coefficient between the two gradings was only 0.31 (Table 5).

Similar findings are observed for the RPE outcomes on intact area. In particular, the ICC is moderate/poor for the central subfield (0.61) and the inner ring (0.62). The ICC is high for the outer

Table 1. Demographic Characteristics of the 30 ProgStar Participants (Eyes) Used in the Repeatability Analysis and Characteristics of the Whole ProgStar Cohort

Demographic in the Reproducibility Sample (N = 30)		Demographic in the Whole ProgStar Cohort (N = 259)	
Sex	*n* (%)	Sex	*n* (%)
Female	14 (46.7)	Female	141 (54.4)
Race		Race	
White	26 (86.7)	White	222 (85.7)
Black	1 (3.3)	Black	20 (7.7)
Asian	2 (6.7)	Asian	10 (3.9)
Other	1 (3.3)	Other or don't know	7 (2.7)

Table 2. Additional Demographic Characteristics of the 30 ProgStar Participants (Eyes) Used in the Repeatability Analysis and Characteristics of the Whole ProgStar Cohort

Repeatability Analysis Demographic in the Reproducibility Sample (N = 30)			Demographic in the Whole ProgStar Cohort (N = 259)		
	Median (IQR)	Range		Median (IQR)	Range
Age (years)	32 (27, 40)	(16–53)	Age (years)	31 (21, 44)	(7–69)
Age of symptom onset (years)	22.5 (17, 29)	(8–48)	Age of symptom onset (years)	19 (12, 29)	(4–64)
Duration of symptoms (years)	8 (5, 12.5)	(1–23)	Duration of symptoms (years)	9 (5, 15)	(0–55)
BCVA (LogMAR)	0.78 (0.60, 0.88)	(0.08–1.16)	BCVA (LogMAR)	0.88 (0.66, 1.00)	(−0.06–1.30)

ring (0.96), although the B-A plot may indicate a pattern for between-grading difference where the measurements from the second grading were often smaller than those from the first grading.

Retinal Layer: OS

Figure 3B shows the B-A plots for the OS. In the ETDRS central subfield, measurements of the thickness/intact area for most of the sample (70%) was 0 (Table 3). The corresponding Kappa statistic on dichotomized thickness was 0.84 (Table 5). In the inner and outer ring regions (Table 4), the P values comparing the mean thickness of the two gradings were both <0.1, with RADs of 21% and 9%, respectively, suggesting potential systematic difference in grading thickness for this layer.

For the outcome on intact area of the OS, it was 0 for most of the sample in the ETDRS central subfield. In the inner and outer ring regions, the ICCs are excellent (0.99 and 0.97, respectively; Table 4). However, the RAD for the inner ring is 9%. The RAD for the outer ring is small (1%), but there is

large variability of the intact area measurements caused by a few outlying points (Fig. 3B).

Retinal Layer: IS

Figure 3C shows the B-A plots for the IS. In the ETDRS central subfield, 73% measurements of the thickness/intact area were 0 (Table 3), and the Kappa statistic on dichotomized thickness was 0.92 (Table 5). In the inner ring, there was an excellent ICC (0.99; Table 4) and no sign of systemic difference between gradings, although RAD was 11%, reflecting the level of between-grading difference (Table 4). In the outer ring, the ICC was 0.89, and there were signs of systematic difference on grading the thickness measure (Fig. 3C; P value from paired t-test = 0.12).

For the outcome on intact area in IS, most of the measurements in the ETDRS central subfield were 0. The inner ring had excellent ICC (0.99) and no sign of systematic difference (Table 4), and the level of difference was low (RAD = 4%). In the outer ring, despite the good ICC (0.92), however, there may be systematic difference most likely driven by a few

Table 3. Summary Statistics of the OCT Measurements of the 30 ProgStar Images Used in the Repeatability Analysis

Layer	Number (%) of Thickness = 0 for Center[a]	Thickness (μm) Mean (SD)			Intact Area (mm²) Mean (SD)		
		Center[a]	Inner Ring[a]	Outer Ring[a]	Center[a]	Inner Ring[a]	Outer Ring[a]
RPE	10 (33.3)	9.28 (11.32)	19.31 (8.01)	27.79 (4.82)	0.32 (0.34)	4.91 (1.45)	20.43 (1.60)
Outer segment	21 (70.0)	2.18 (11.32)	7.04 (7.07)	16.78 (5.27)	0.11 (0.23)	2.87 (2.00)	19.45 (2.68)
Inner segment	20 (66.7)	3.78 (8.43)	11.30 (9.81)	26.42 (5.22)	0.12 (0.24)	2.95 (2.01)	19.63 (2.58)
Outer nuclear layer	1 (3.3)	30.77 (25.71)	54.22 (12.21)	69.44 (8.73)	0.54 (0.29)	6.10 (0.39)	20.78 (0.33)
Inner retina	0	78.17 (17.02)	142.77 (11.37)	130.95 (8.57)	0.78 (0.03)	6.28 (0.02)	20.78 (0.33)
Total retina	0	148.13 (39.25)	248.13 (29.89)	274.36 (21.54)	0.79 (0.01)	6.28 (0.02)	20.78 (0.33)

[a] The summary statistics used OCT data from the original grading of the ProgStar images (i.e. grading 1).
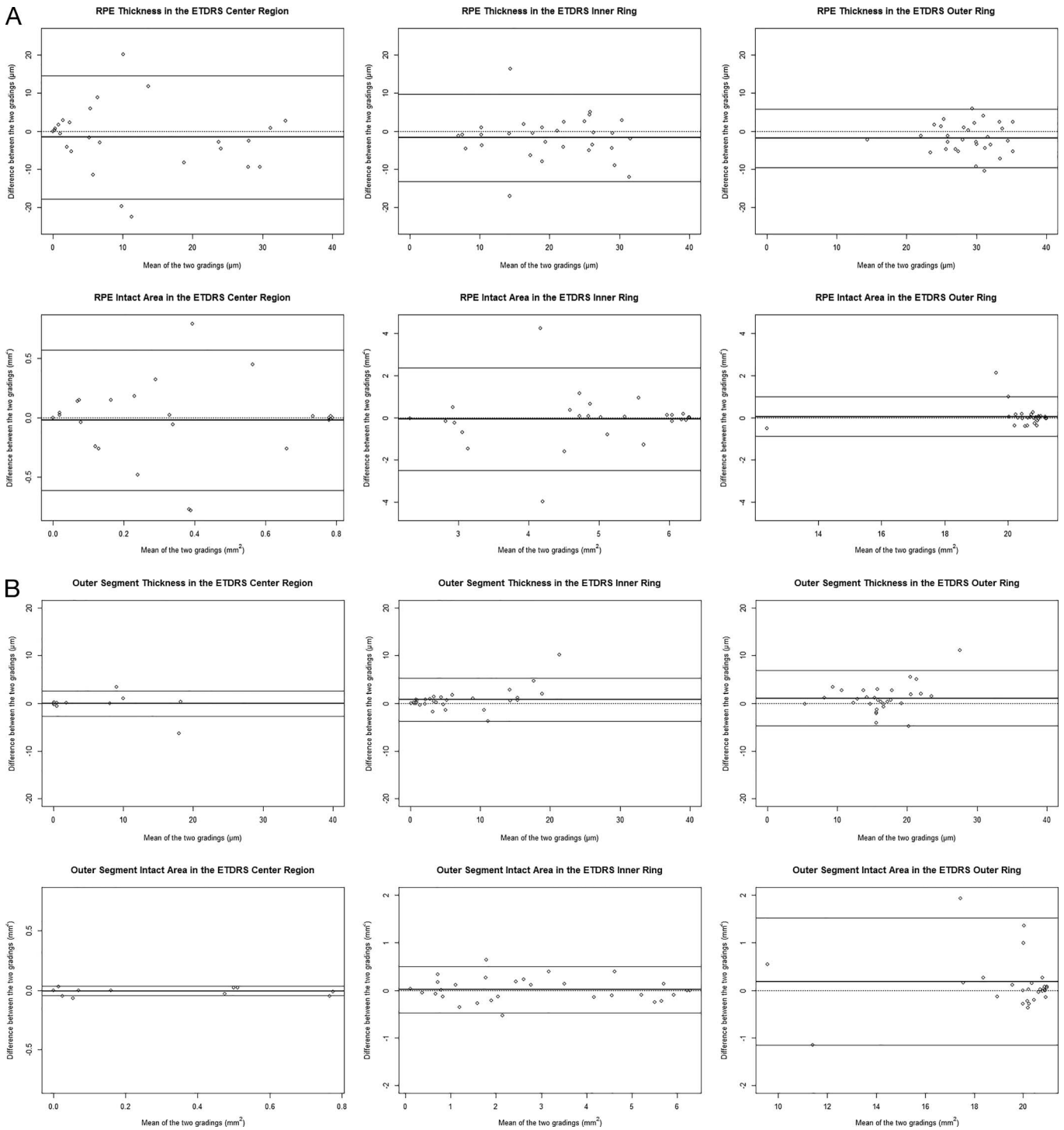
translational vision science & technology

Figure 3. Bland-Altman Plots for thickness and intact area of retinal layers of (A) RPE, (B) outer segment, (C) inner segment, (D) outer nuclear layer, (E) inner retina, and (F) the total retina in the ETDRS center and inner ring and outer ring regions. In each plot, the *darkened line* is the mean of the difference of two gradings (grading 2 – grading 1), the *lighter lines* are the 95% upper and lower limits of the mean difference, and the *dotted line* is for difference = 0.
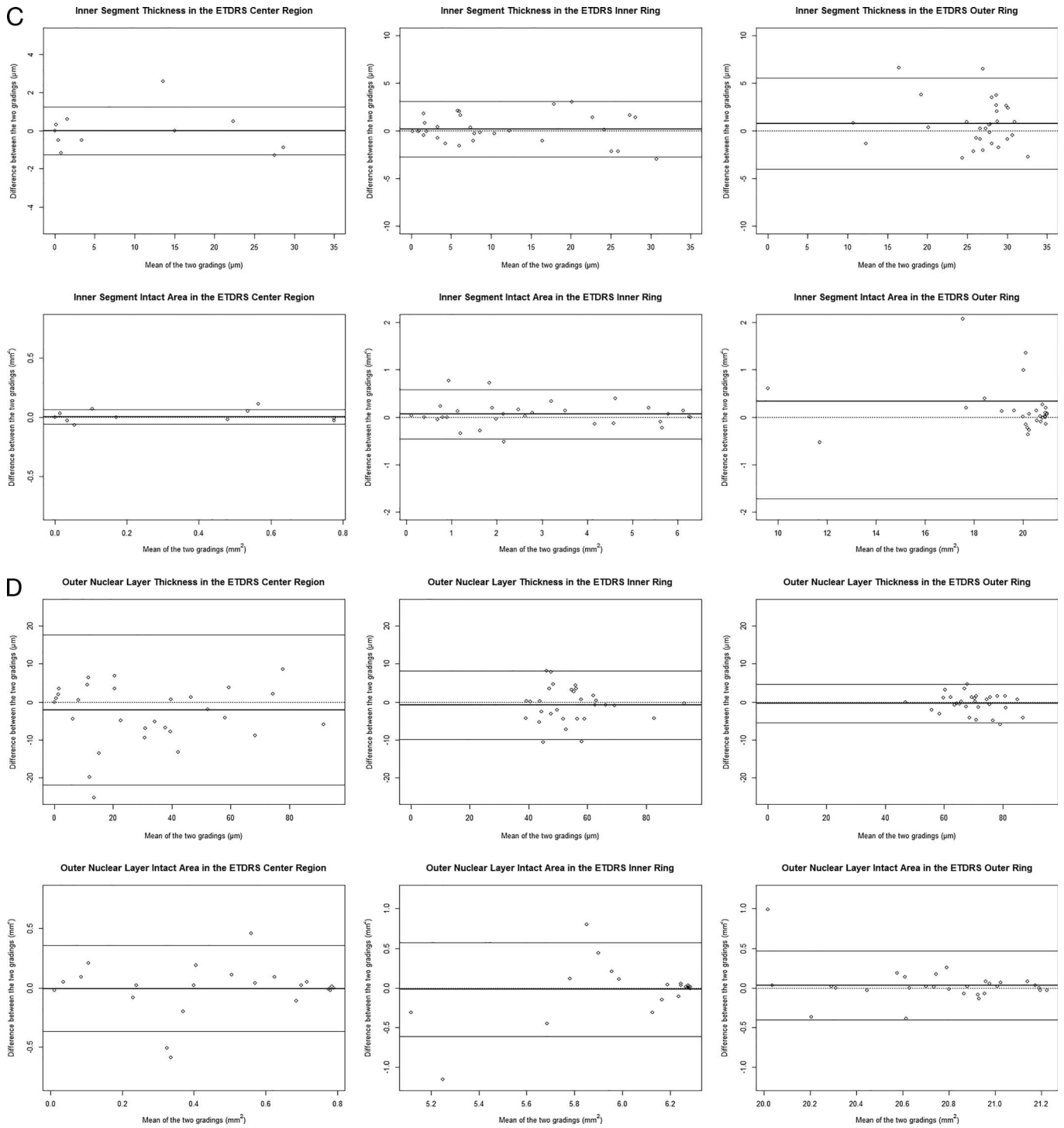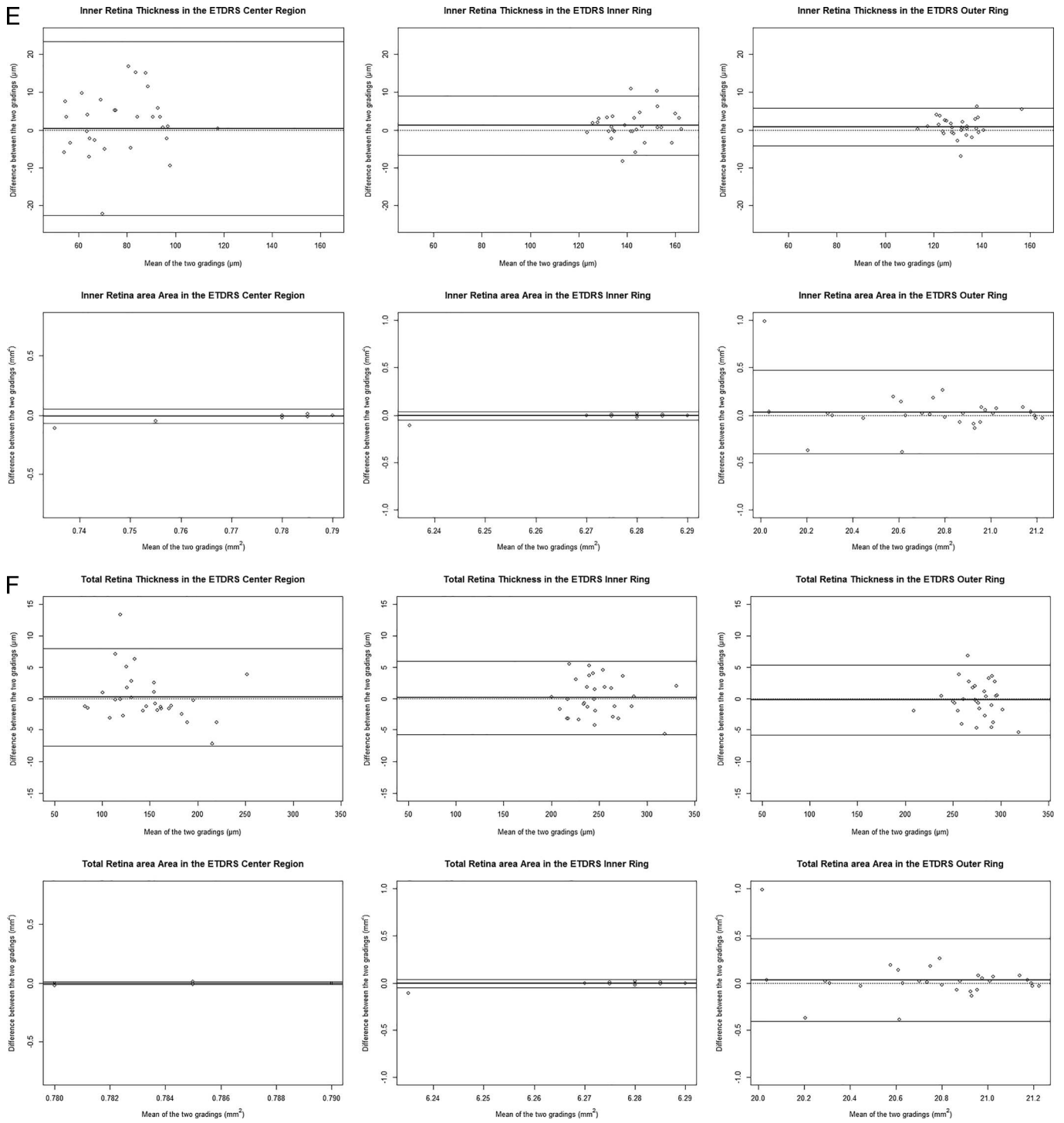
C

D

Figure 3. Continued.

E

Inner Retina Thickness in the ETDRS Center Region

Inner Retina Thickness in the ETDRS Inner Ring

Inner Retina Thickness in the ETDRS Outer Ring

Inner Retina area Area in the ETDRS Center Region

Inner Retina area Area in the ETDRS Inner Ring

Inner Retina area Area in the ETDRS Outer Ring

F

Total Retina Thickness in the ETDRS Center Region

Total Retina Thickness in the ETDRS Inner Ring

Total Retina Thickness in the ETDRS Outer Ring

Total Retina area Area in the ETDRS Center Region

Total Retina area Area in the ETDRS Inner Ring

Total Retina area Area in the ETDRS Outer Ring

Figure 3.   Continued.

translational vision science & technology

Table 4. Repeatability Assessments Based on the Paired *t*-test, Absolute Difference, and ICC Between the Two Gradings for All Retina Layers in the Center, Inner Ring, and Outer Ring

Layer		Thickness (μm)		
		Center	Inner Ring	Outer Ring
RPE	Paired *t*-test comparing means of the two gradings, *P* value	0.30	0.13	0.013
	Absolute difference between two gradings, median (IQR)	2.90 (0.70, 8.90)	2.85 (1.00, 5.00)	2.80 (1.70, 4.80)
	Relative absolute difference median (IQR) (%)[a]	NE[†]	0.14 (0.09, 0.28)	0.11 (0.07, 0.17)
	ICC	0.76	0.73	0.63
Outer segment	*P* value from paired *t*-test comparing means of the two gradings	0.76	0.09	0.06
	Absolute difference between two gradings, median (IQR)	0 (0,0.10)	0.80 (0.30,1.50)	1.40 (0.70,2.70)
	Relative absolute difference median[a] (IQR)	NE[†]	0.21 (0.08,0 .34)	0.09 (0.04, 0.19)
	ICC	0.97	0.93	0.79
Inner segment	Paired *t*-test comparing means of the two gradings, *P* value	0.91	0.60	0.12
	Absolute difference between two gradings, median (IQR)	0 (0, 0.50)	1.10 (0.30, 1.80)	1.40 (0.80, 2.70)
	Relative absolute difference median[a] (IQR)	NE[†]	0.11 (0.05, 0.26)	0.07 (0.03, 0.09)
	ICC	0.99	0.99	0.89
Outer nuclear layer	Paired *t*-test comparing means of the two gradings, *P* value	0.26	0.33	0.38
	Absolute difference between two gradings, median (IQR)	5.05 (2.10, 8.60)	3.3 (0.80, 4.50)	1.35 (0.70, 3.20)
	Relative absolute difference median[a] (IQR)	0.22 (0.08, 0.84)	0.06 (0.01, 0.09)	0.02 (0.01, 0.05)
	ICC	0.92	0.93	0.96
Inner retina	Paired *t*-test comparing means of the two gradings, *P* value	0.88	0.11	0.10
	Absolute difference between two gradings, median (IQR)	5.15 (3.40, 9.40)	2.1 (0.60, 3.60)	1.20 (0.60, 2.80)
	Relative absolute difference median[a] (IQR)	0.07 (0.04, 0.13)	0.02 (0, 0.03)	0.01 (0.01, 0.02)
	ICC	0.77	0.93	0.95
Total retina	Paired *t*-test comparing means of the two gradings, *P* value	0.77	0.84	0.64
	Absolute difference between two gradings, median (IQR)	1.80 (1.20, 3.80)	2.00 (1.30, 3.60)	1.90 (0.50, 3.60)
	Relative absolute difference median[a] (IQR)	0.01 (0.01, 0.02)	0.01 (0.01, 0.01)	0.01 (0, 0.01)
	ICC	0.99	0.99	0.99

[a] Relative absolute difference (%) is calculated as the median of (the absolute difference between two gradings /by the average of the two gradings), that is, it reflects the level of difference between gradings compared to the magnitude of the measurement of the variable. NE[†], not estimable if the denominator is 0, that is, average of two gradings for the layer is 0; NE, ICC not estimable because there is minimal variability in the data values (i.e. the denominator for ICC is close to 0).

Table 4. Extended

Layer		Center	Inner Ring	Outer Ring
			Intact Area (mm²)	
RPE	Paired *t*-test comparing means of the two gradings, *P* value	0.71	0.74	0.58
	Absolute difference between two gradings, median (IQR)	0.04 (0.00, 0.26)	0.17 (0.04, 0.81)	0.07 (0.03, 0.28)
	Relative absolute difference median (IQR) (%)[a]	NE[†]	0.03 (0.01, 0.17)	0 (0, 0.01)
	ICC	0.61	0.62	0.96
Outer segment	*P* value from paired *t*-test comparing means of the two gradings	0.22	0.77	0.15
	Absolute difference between two gradings, median (IQR)	0 (0,0.01)	0.14 (0.10,0.27)	0.15 (0.03,0.29)
	Relative absolute difference median[a] (IQR)	NE[†]	0.09 (0.04, 0.16)	0.01 (0, 0.01)
	ICC	0.99	0.99	0.97
Inner segment	Paired *t*-test comparing means of the two gradings, *P* value	0.56	0.23	0.09
	Absolute difference between two gradings, median (IQR)	0 (0, 0.02)	0.13 (0.04, 0.23)	0.14 (0.06,0 .37)
	Relative absolute difference median[a] (IQR)	NE[†]	0.04 (0.02, 0.11)	0.01 (0, 0.02)
	ICC	1.00	0.99	0.92
Outer nuclear layer	Paired *t*-test comparing means of the two gradings, *P* value	0.88	0.72	0.44
	Absolute difference between two gradings, median (IQR)	0.02 (0.01, 0.11)	0.02 (0.01, 0.15)	0.04 (0.02, 0.14)
	Relative absolute difference median[a] (IQR)	0.07 (0.01, 0.52)	0 (0, 0.02)	0 (0, 0.01)
	ICC	0.79	0.61	0.80
Inner retina	Paired *t*-test comparing means of the two gradings, *P* value	0.10	0.41	0.44
	Absolute difference between two gradings, median (IQR)	0 (0, 0.01)	0 (0, 0.01)	0.04 (0.02, 0.14)
	Relative absolute difference median[a] (IQR)	0 (0, 0.01)	0 (0, 0)	0 (0, 0.01)
	ICC	NE	NE	0.80
Total retina	Paired *t*-test comparing means of the two gradings, *P* value	1.00	0.41	0.44
	Absolute difference between two gradings, median (IQR)	0 (0, 0.00)	0 (0, 0.01)	0.04 (0.02, 0.14)
	Relative absolute difference median[a] (IQR)	0 (0, 0.01)	0 (0, 0)	0 (0, 0.01)
	ICC	0.42	NE	0.80

samples with lower intact area but with large grading difference (paired *t*-test *P* value = 0.09; Fig. 3C).

Retinal Layer: ONL

Figure 3D shows the B-A plots for the ONL. For the thickness measure, in the ETDRS central subfield, the ICC was 0.92. Although there was no significant mean difference between gradings based on the paired *t*-test, the level of absolute difference was high, with RAD = 22% (Table 4). Thickness in the inner ring and outer ring had good ICCs (0.93 and 0.96, respectively) and smaller RADs (6% and 2%, respectively).

For intact area, the central subfield had low ICC (0.79). In the inner and outer ring, although the

Table 5. Agreement on Prevalence of Loss of a Layer for the RPE, Inner Segment, and Outer Segment in the ETDRS Central Subfield

Layer	RPE	Outer Segment	Inner Segment	Outer Nuclear Layer	Inner Retina
Agreement on thickness = 0 (n)	5	20	20	0	0
Agreement on thickness > 0 (n)	16	8	9	28	30
Disagreement (n)	9	2	1	2	0
% of Agreement	70	93.3	96.7	93.3	100
Kappa statistic	0.31	0.84	0.92	−0.03	NE*

* NE, not estimated because all had thickness >0.

median RAD was both 0, the ICCs were poor (0.61 and 0.80, respectively), and the B-A plot showed signs of heterogeneous difference where for smaller intact area values, the second grading was often smaller than the first grading (Fig. 3D).

Retinal Layer: Inner Retina

Figure 3E shows the B-A plots for the inner retina. For thickness, there was large between-grading difference for the central subfield, with ICC = 0.77 (Table 4). For the inner and outer ring, the ICCs were good (0.93 and 0.95, respectively), but there was more positive difference comparing grading 2 to grading 1 (Fig. 3E) (paired t-test P values = 0.11 and 0.10 respectively).

For intact area, there was most often no difference between gradings in the central subfield and the inner ring, but measurements of intact area themselves lacked variability (i.e., the intact area measurements focused on few data values). Intact area in the outer ring was more variable, but agreement was poor (ICC = 0.80), and the B-A plot showed signs of heterogeneous difference where for smaller intact area values, the second grading was often smaller than the first grading.

Retinal Layer: Total Retina

Figure 3F are the B-A plots for the total retina. For thickness of all three ETDRS regions, the ICCs were excellent (all 0.99; Table 4), there were no significant mean differences between gradings, and the RADs were minimal. For the central subfield, however, the B-A plot shows signs of heterogeneous difference where for larger thickness values, the second grading was often smaller than the first grading.

For the outcome of intact area, for the central subfield and inner ring, the data focused on few values and, thus, there was minimal variability in grading

difference. For the outer ring, the agreement was poor (ICC = 0.8), and the B-A plot shows signs of heterogeneous difference and heteroscedastic variability.

Discussion

Quantitative structural parameters from SD-OCT are important outcomes in the ProgStar study and are important candidates of endpoints for future treatment trials of STGD1. Using 30 randomly selected images from the ProgStar cohort, we assessed the reproducibility of measurements of the SD-OCT parameters, including the thickness and intact area of six retinal layers in three ETDRS regions covering the fovea and parafoveal areas.

The high prevalence of complete loss of the IS and OS in the central subfield is consistent with the pathology of STGD1 where photoreceptors are primarily affected and corresponds to visual dysfunction.[1] The level of intactness of the EZ band is of particular scientific and regulatory interest.[16] However, in this cohort, we observed highly heterogeneous and sporadic partial disruption of the EZ band. Combined with the disorganization of the outer retina, it became very challenging to accurately segment the relevant boundaries and categorize tissues into the appropriate layers (Figs. 4–6). In the ETDRS central subfield, given the high prevalence of 0 thickness (and intact area) of IS and OS, the thickness and intact area outcomes may be dichotomized for analysis (=0 vs. >0), although the Kappa statistic only showed moderate agreement on the dichotomized assessments. In the outer ring, the thickness or intact area of the IS or OS either had low ICC, indicating a potentially systematic difference between grading, or a pattern of data distribution in the B-A plots suggesting heteroscedastic measurement error for the measure and, thus, may not provide
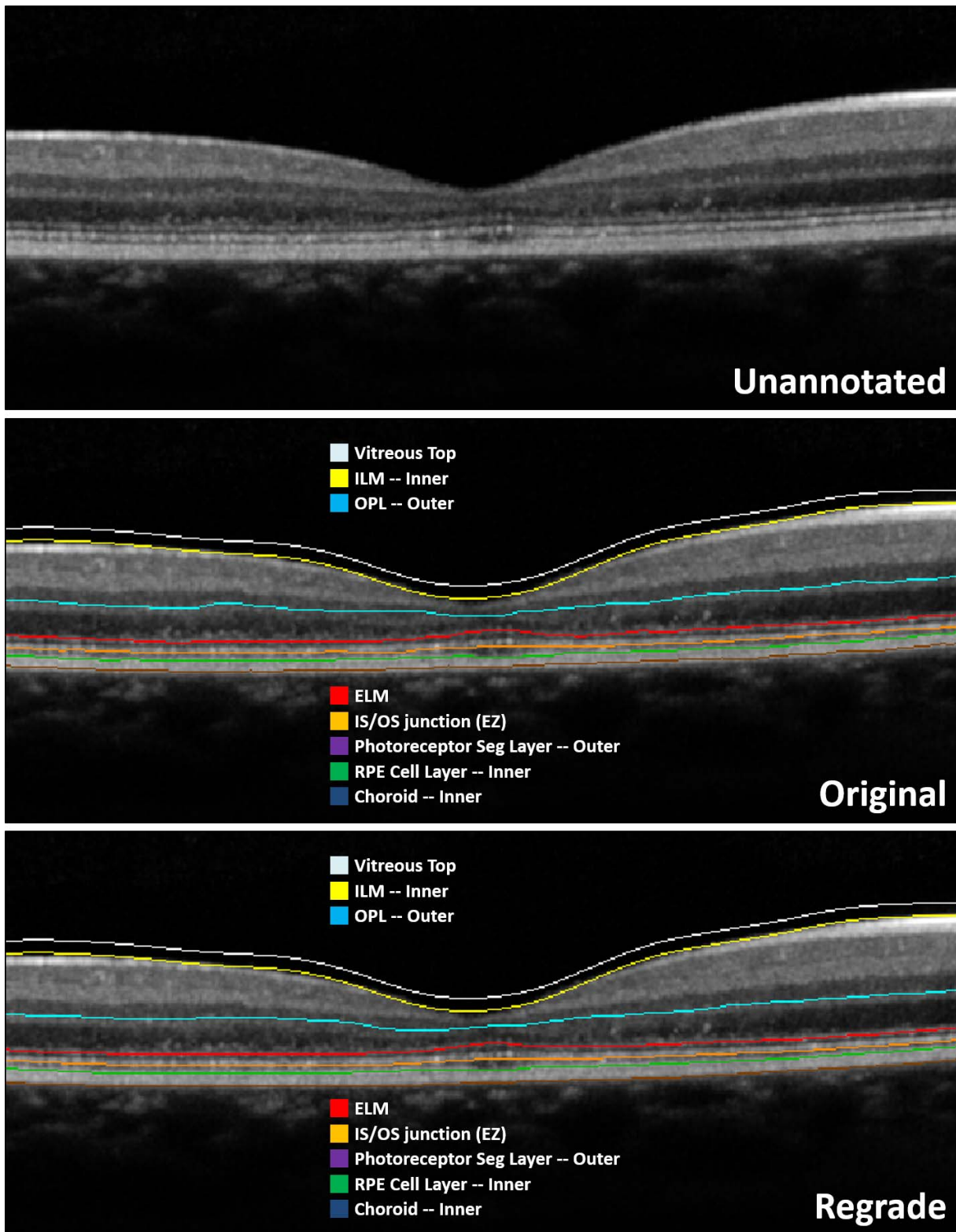
Figure 4. An example of an unambiguous case in the ProgStar reproducibility study. Screenshots are provided showing the unannotated image and the segmentation from the original grading and regrading, respectively. Agreement on segmentation between the two gradings was good because the morphology was very clear.
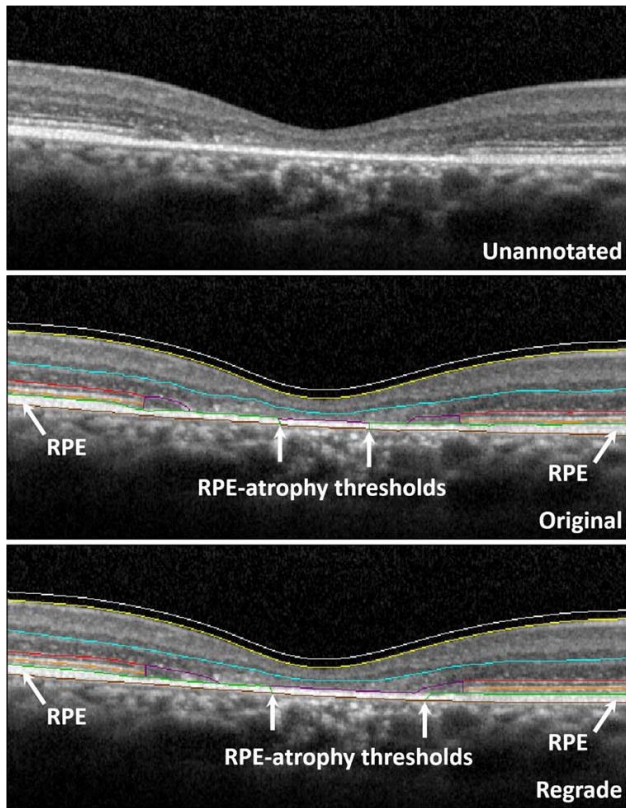
Figure 5. Example of an ambiguous case in the ProgStar reproducibility study where determination of the RPE-atrophy transition thresholds was different in the two gradings. Screenshots are provided showing the unannotated image and the segmentation from the original grading and regrading, respectively. Note RPE is the layer between the *green* and *brown lines*. RPE absence is denoted in the central A-scans where the *green line* is absent, as it is overlapping with the *brown line* corresponding with Bruch's membrane (the two lines are snapped together, representing zero thickness).
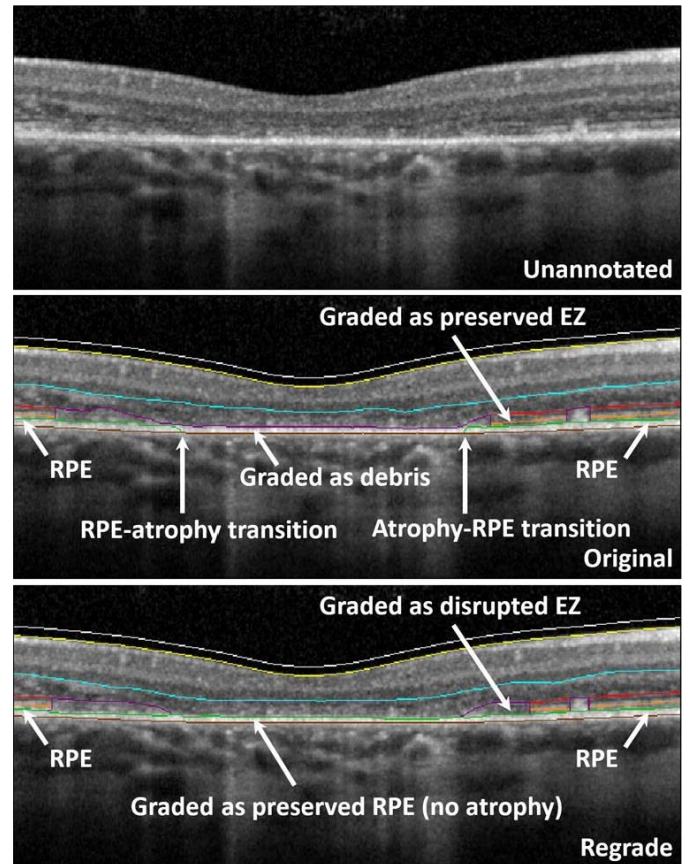

Figure 6. An example of an ambiguous case in the ProgStar reproducibility study where determination of preserved RPE versus debris was different between gradings. Screenshots are provided showing the unannotated image and the segmentation from the original grading and regrading, respectively. Note RPE is the layer between the *green* and *brown lines*. RPE absence is denoted in the original grading in the central A-scans where the *green line* is absent, as it is overlapping with the *brown line* corresponding with Bruch's membrane (the two lines are snapped together, representing zero thickness). There is also slight disagreement in EZ preservation.

reliable assessments of IS and OS in the outer ring region. In the inner ring, thicknesses of IS and OS had large absolute between grading difference (median RADs > 10%). In future studies, if photoreceptor segment thickness measures are of interest, agglomerating the IS and OS thicknesses into a single layer may be more useful than the individual layer thicknesses because of the poor reliability as observed in our data.

On the other hand, the outcome of the intact areas of IS and OS in the inner ring had excellent ICC, smaller between grading difference, and no strong patterns in the B-A plots. Comparable results on intergrader reproducibility of the area of EZ loss in the total SD-OCT macular scans were also observed in another study.[17] Therefore, the outcome of intact area of the IS and OS layers may provide sensible measurements of the EZ band in STGD1.

Measurements for the RPE generally had low ICC or showed some pattern in the B-A plots, indicating heteroscedastic measurement error, suggesting that reliably measuring thickness or intact area of RPE can be difficult. Presentation of the disease was highly heterogeneous, with some eyes having clear morphological features similar to atrophy in dry age-related macular degeneration (AMD) (Fig. 4), whereas other eyes having ambiguous features due to a gradual transition zone (Figs. 5, 6). Therefore, it was very challenging to determine the exact A-scan or threshold at which the RPE transitioned from present to

absent. Despite that heuristics for establishing atrophy presence, like those suggested for geographic atrophy,[18] were adapted in grading the ProgStar images, we could not determine the true nature and precise cellular processes represented by the hyper-reflective features (e.g., flecks, debris, and compromised RPE) and found that the transition zones were often too gradual (Figs. 5, 6). The ambiguous presence of transition zones may also explain the systematic differences between gradings for some parameters, where one grader's preference in marking the involved boundaries may be often different from the other grader (Fig. 6).

Our data show that measurements of the RPE layer had considerable noise. However, these measurements may still have utility in quantifying structural changes in the specific context involving STGD1 patients with a specific phenotype (i.e., unambiguous atrophy like in dry AMD). Nonetheless, such changes will not be generalizable to the highly heterogeneous general STGD1 population. A potential area of exploration is the use of alternative, voxel intensity-based segmentation methods, such as polarization sensitive OCT,[19] which could be used to supplement a grader's determination of the precise point of RPE atrophy.

The thickness measure for ONL in the ETDRS central subfield had large absolute difference between gradings, but thicknesses in the inner ring and outer ring may provide reliable assessments of this layer. The intact areas of ONL in all three ETDRS regions had low ICCs and will not be efficient (i.e., not powerful) if used as endpoints for future trials.

The inner retina and the total retina are thicker layers. Intact areas of these two layers in all three ETDRS regions all had very small variability of the data points or low ICC, and thus, the measure of intact area would not be a sensitive measure to capture structural changes of the inner retina and the total retina. This is in line with the pathophysiology of STGD1 where little, if any, change is anticipated to affect the inner retina. Thickness of the inner retina in all three ETDRS regions had low ICC or signs of systematic difference between gradings and, thus, may not provide unbiased estimate of disease progression in this layer. Thickness of the total retina in the ETDRS central subfield showed potential heteroce-dastic measurement error in the B-A plot (i.e., difference between gradings depended on the magnitude of the gradings), which may generate bias when estimating rate of thickness loss by using this measure. Thicknesses of the total retina in the inner
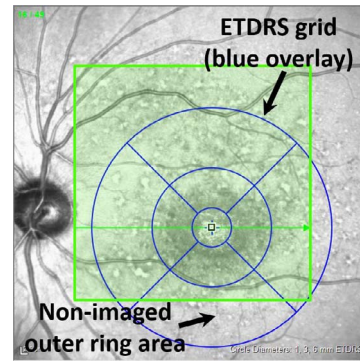

Figure 7. An example of a scan with incomplete scan area that made the ETDRS outer ring region not fully measured. The *green covered area* is the scanned area, whereas the *blue circles* are the ETDRS grid.

and outer rings have excellent ICCs and small relative absolute difference and, thus, are good candidates for tracking disease progression and for serving as endpoints for future clinical trials.

Across all layers, measurements of the ETDRS central subfield consistently had low ICC, large absolute difference between gradings, and/or lack of variability of the data points. For the thinner layers of RPE, OS, and IS, even the agreement of dichotomized thickness (0 vs. >0) was poor to moderate. This suggests that precise measurements of the foveal region in STGD were challenging. This finding was similar to what has been reported for neovascular AMD.[20] The finding could also be explained in consideration of ProgStar's enrollment criteria where eyes were enrolled only if they already showed area of atrophy in fundus photos. At this stage of disease pathogenesis, accurate segmentation of the foveal central region can be difficult (Figs. 5, 6). In particular, in the central subfield, most eyes had IS thickness and, hence, intact area measured as 0, suggesting total EZ disruption in the foveal region in these eyes. Thus, measures of the EZ in the central subfield cannot be used to track further disease progression.

For measurements of the ETDRS outer ring, irrespective of the reliability performance assessed here, they are subject to another source of measurement error due to incomplete scan coverage of the ETDRS grid (Fig. 7). The ProgStar study protocol required that photographers center the cube scan at the anatomic fovea and if it was a follow-up visit use the follow-up and eye-tracking function of the device. However, exactly locating the fovea could be difficult for the photographers because ProgStar participants, by enrollment criteria, already had a sizable macula

atrophy and most of them had eccentric fixation. During image reading, graders determined the fovea center by using B-scans and then overlaid the ETDRS grid. Thus, the center used by photographers during image acquisition could be different from the center determined during grading, and consequently, the ETDRS outer ring region may not be fully captured in the scan. Additionally, at a follow-up visit, some patients may have acquired a new preferred retinal locus, and the "follow-up" and "eye tracking" options of the Heidelberg device could not apply. The image then had to be taken using the "baseline" option. Therefore, for an individual eye, the actual region of the retina covered in the total scan may not be identical over different visits. These issues may lead to artifacts in measurements of the outer ring region and the total $20° \times 20°$ scan area. Therefore, measurements of the outer ring and also the total scan area may not be good candidates for trial endpoints. It is also important that any downstream analysis of these measurements should consider and address the measurement error associated with these artifacts.

SD-OCT has been widely used by researchers and clinicians for quantitative assessment of pathologies, and high-quality scans have been shown to be obtainable from clinical or research applications.[21] Excellent repeatability and reproducibility of retinal layer thickness measurements have also been demonstrated in healthy eyes using Spectralis SD-OCT.[22–24] However, reliability of quantitative SD-OCT measurements is more variable in eyes with pathologies and depends on the accuracy of segmentation of retinal layers. Repeatability studies of eyes with diabetic macular edema[25] and neovascular AMD[20] did not identify segmentation errors as a major problem. However, limitations were noted of using SD-OCT-derived structural measurements as endpoints in glaucoma trials because of the variability due to segmentation error and discrepancy between SD-OCT devices and other sources.[26–28] Increased automatic segmentation error was also noted in pathologies such as AMD,[20,29,30] and a high algorithm error rate (34%) had been reported for SD-OCT scans in STGD1 macular dystrophy.[31]

Limitations of this study include that we evaluated reproducibility[9] of grading of SD-OCT parameters (i.e., repeated grading of the same scans but by different graders), but the repeatability[9] associated with repeated SD-OCT scans was not assessed. Also, although there was dual-grader evaluation in ProgStar with each image, the two graders were not masked to each other; thus, assessing grader specific variability was not possible. Instead, this study assessed the reproducibility of the grading process adopted in ProgStar. Another limitation is that there were no control measurements from corresponding OCT data from age-matched normal eyes. A prior study from the DIRC compared preliminary OCT grading results for a subset of ProgStar baseline visit images to a convenience sample of 20 normal eyes by using the ProgStar grading protocol and found that the thicknesses of all the outer retina layers of the STGD1 eyes were all statistically significantly thinner than normal eyes (Ho A, et al. *IOVS*. 2016;57:ARVO E-Abstract 2697). However, the clinical significance of the retinal thinning compared to normal eyes was not evaluated.

In summary, in ProgStar SD-OCT grading, the complex morphology and confounding pathology (e.g., intraretinal cystoid spaces, outer retinal tubulation, hyperreflective flecks and debris fields, ambiguous RPE atrophy, and collapse and disruption of intraretinal layers overlying compromised RPE) of the outer retina especially in the foveal region made automatic segmentation impractical, and significant effort was needed to conduct manual segmentation, which limited the use of the detailed SD-OCT grading scheme adopted in the ProgStar study for measuring structural end-points for future STGD1 trials. This also resulted in difficulties to reliably measure thickness or intact area of the retinal layers in the central subfield. For the outer ring, the measures of intact area either had poor reproducibility or limited variability, and these measures particularly suffered from the potential biases due to incomplete scan coverage of the ETDRS grid. If assuming that the thicknesses measured on the scanned outer ring region truly reflected the complete outer ring region, the thicknesses of the ONL and the total retina may be used to evaluate disease progression in the outer ring region in STGD1. Measurements for the total scan area ($20° \times 20°$) were not evaluated here. However, as aforementioned, they would not be suitable for tracking disease progression because of the possible change in scan-covered retinal region.

For STGD1 eyes that already had identifiable atrophy as those in ProgStar, measurements in the inner ring, including IS intact area, ONL thickness, and the total retina thickness, had good reproducibility. The much-reduced intact areas in the layers of IS, OS, and RPE in the ETDRS inner ring region as compared to the theoretical area of this region suggest that retinal damage was observed and quantifiable using the intact area measure in these layers in this

parafoveal region. Thickness of the total retina in this region, thus, must also have been impacted by the disease. ONL thickness was previously shown to be significantly thinner than normal eyes (Ho A, et al. *IOVS*. 2016;57:ARVO E-Abstract 2697). Therefore, taken together, the reproducibility performance and reflection of disease progression, the measures of IS intact area, ONL thickness, and total retina thickness in the inner ring region may provide promising measures that are both reliable and sensitive to show change. In particular, the measure of intact area for IS can reflect the status of the EZ, which has been of regulatory agencies' focus as a potential endpoint for geographic atrophy and inherited retinal degenerations.[32] Our data suggest that in ProgStar, most eyes had complete EZ loss in the ETDRS central subfield already. Damage of EZ in the ETDRS inner ring region was observed, and the preservation of the EZ in this region can be reliably measured. Therefore, integrity of EZ in the ETDRS inner ring region holds good potential as a structural endpoint for STGD1 trials, and segmentation effort should specifically focus on the EZ related layers, that is, IS and OS, involving the ELM, IS/OS junction, and the outer photoreceptor segment layer. However, to determine the ultimate suitability of the measure of EZ integrity (i.e., IS intact area) and the measure of ONL and total retinal thickness in the inner ring as endpoints for future treatment trials, structure-function analysis is necessary to assess cross-sectional and longitudinal relationships of these measures with visual function loss.

References

1. Tanna P, Strauss RW, Fujinami K, Michaelides M. Stargardt disease: clinical features, molecular genetics, animal models and therapeutic options. *Br J Ophthalmol*. 2017;101:25–30.

2. Dalkara D, Goureau O, Marazova K, Sahel JA. Let there be light: gene and cell therapy for blindness. *Hum Gene Ther*. 2016;27:134–147.

3. Schwartz SD, Regillo CD, Lam BL, et al. Human embryonic stem cell-derived retinal pigment epithelium in patients with age-related macular degeneration and Stargardt's macular dystrophy: follow-up of two open-label phase 1/2 studies. *Lancet*. 2015;385:509–516.

4. Thompson DA, Ali RR, Banin E, et al. Advancing therapeutic strategies for inherited retinal degeneration: recommendations from the Monaciano Symposium. *Invest Ophthalmol Vis Sci*. 2015;56:918–931.

5. Scholl HP, Strauss RW, Singh MS, et al. Emerging therapies for inherited retinal degeneration. *Sci Transl Med*. 2016;8:368rv6.

6. Strauss RW, Ho A, Munoz B, et al. The natural history of the progression of atrophy secondary to Stargardt disease (ProgStar) studies: design and baseline characteristics: ProgStar Report No. 1. *Ophthalmology*. 2016;123:817–828.

7. Cukras C, Jeffrey BG. The importance of outcome measure research in Stargardt disease. *JAMA Ophthalmol*. 2017;135:704–705.

8. Al-Mujaini A, Wali UK, Azeem S. Optical coherence tomography: clinical applications in medical practice. *Oman Med J*. 2013;28:86–91.

9. Bartlett JW, Frost C. Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. *Ultrasound Obstet Gynecol* 2008;31:466–475.

10. Keane PA, Liakopoulos S, Ongchin SC, et al. Quantitative subanalysis of optical coherence tomography after treatment with ranibizumab for neovascular age-related macular degeneration. *Invest Ophthalmol Vis Sci*. 2008;49:3115–3120.

translational vision science & technology

11. Lujan BJ, Roorda A, Croskrey JA, et al. Directional optical coherence tomography provides accurate outer nuclear layer and Henle fiber layer measurements. *Retina*. 2015;35:1511–1520.

12. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Med*. 1999;8:135–160.

13. SAS Product Documentation. 1995. https://support.sas.com/documentation/onlinedoc/stat/ex_code/131/intracc.html.

14. McHugh ML. Interrater reliability: the kappa statistic. *Biochemia Med*. 2012;22:276–282.

15. SAS INSTITUTE INC. SAS intracc: Intraclass Correlations Available at: https://support.sas.com/documentation/onlinedoc/stat/ex_code/131/intracc.html.

16. Csaky K, Ferris F, 3rd, Chew EY, Nair P, Cheetham JK, Duncan JL. Report from the NEI/FDA endpoints workshop on age-related macular degeneration and inherited retinal diseases. *Invest Ophthalmol Vis Sci*. 2017;58:3456–3463.

17. Cai CX, Light JG, Handa JT. Quantifying the rate of ellipsoid zone loss in Stargardt disease. *Am J Ophthalmol*. 2018;186:1–9.

18. 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;125:537–548.

19. Ritter M, Zotter S, Schmidt WM, et al. Characterization of stargardt disease using polarization-sensitive optical coherence tomography and fundus autofluorescence imaging. *Invest Ophthalmol Vis Sci*. 2013;54:6416–6425.

20. Patel PJ, Chen FK, Ikeji F, et al. Repeatability of stratus optical coherence tomography measures in neovascular age-related macular degeneration. *Invest Ophthalmol Vis Sci*. 2008;49:1084–1088.

21. McLellan GJ, Rasmussen CA. Optical coherence tomography for the evaluation of retinal and optic nerve morphology in animal subjects: practical considerations. *Vet Ophthalmol*. 2012;15:13–28.

22. Ctori I, Huntjens B. Repeatability of foveal measurements using spectralis optical coherence tomography segmentation software. *PLoS One*. 2015;10:e0129005.

23. Wolf-Schnurrbusch UE, Ceklic L, Brinkmann CK, et al. Macular thickness measurements in healthy eyes using six different optical coherence tomography instruments. *Invest Ophthalmol Vis Sci*. 2009;50:3432–3437.

24. Pinilla I, Garcia-Martin E, Fernandez-Larripa S, Fuentes-Broto L, Sanchez-Cano AI, Abecia E. Reproducibility and repeatability of Cirrus and Spectralis Fourier-domain optical coherence tomography of healthy and epiretinal membrane eyes. *Retina*. 2013;33:1448–1455.

25. Comyn O, Heng LZ, Ikeji F, et al. Repeatability of Spectralis OCT measurements of macular thickness and volume in diabetic macular edema. *Invest Ophthalmol Vis Sci*. 2012;53:7754–7759.

26. Weinreb RN, Kaufman PL. Glaucoma research community and FDA look to the future, II: NEI/FDA Glaucoma Clinical Trial Design and Endpoints Symposium: measures of structural change and visual function. *Invest Ophthalmol Vis Sci*. 2011;52:7842–7851.

27. Krebs I, Smretschnig E, Moussa S, Brannath W, Womastek I, Binder S. Quality and reproducibility of retinal thickness measurements in two spectral-domain optical coherence tomography machines. *Invest Ophthalmol Vis Sci*. 2011;52:6925–6933.

28. Krebs I, Hagen S, Smretschnig E, Womastek I, Brannath W, Binder S. Reproducibility of segmentation error correction in age-related macular degeneration: Stratus versus Cirrus OCT. *Br J Ophthalmol*. 2012;96:271–275.

29. Fiore T, Androudi S, Iaccheri B, et al. Repeatability and reproducibility of retinal thickness measurements in diabetic patients with spectral domain optical coherence tomography. *Curr Eye Res*. 2013;38:674–679.

30. Patel PJ, Chen FK, Ikeji F, Tufail A. Intersession repeatability of optical coherence tomography measures of retinal thickness in early age-related macular degeneration. *Acta Ophthalmologica*. 2011;89:229–234.

31. Strauss RW, Munoz B, Wolfson Y, et al. Assessment of estimated retinal atrophy progression in Stargardt macular dystrophy using spectral-domain optical coherence tomography. *Br J Ophthalmol*. 2016;100:956–962.

32. Csaky KG, Richman EA, Ferris FL, 3rd. Report from the NEI/FDA Ophthalmic Clinical Trial Design and Endpoints Symposium. *Invest Ophthalmol Vis Sci*. 2008;49:479–489.

translational vision science & technology