

Supplementary Online Content

Wilson M, Chopra R, Wilson MZ, et al. Validation and clinical applicability of whole-volume automated segmentation of optical coherence tomography in retinal disease using deep learning. *JAMA Ophthalmol*. Published online July 8, 2021. doi:10.1001/jamaophthalmol.2021.2273

eMethods. Grading Protocol and Disease Severity Grading Definition

eFigure 1. Custom Viewer Used to Assess the 3 Segmentations for the Qualitative Evaluation

eFigure 2. Example Segmentations

eFigure 3. Example 1 of a Scan Where the Model Was Qualitatively Rated Higher Than the 2 Manual Segmentations

eFigure 4. Example 2 of a Scan Where the Model Was Qualitatively Rated Higher Than the 2 Manual Segmentations

eFigure 5. Example of a Scan Where the Model Was Ranked as Most Representative

eFigure 6. Example of a Scan Where All 3 Gradings Were Rated Highly

eFigure 7. Example of a Challenging Case Resulting in Large Disagreement in SRF Volume Between Model and Manual Gradings

eFigure 8. Example of Large Disagreement in SHRM and PED Volume Between Model and Manual Gradings

eFigure 9. Example of Model Failure and Largest Disagreement in PED Volume Between the Model and Expert Graders

eFigure 10. Example of a Model Failure and Large Disagreement in IRF/SRF Volume

eFigure 11. Example 1 of a Disagreement Between Specialists for Qualitative Evaluation

eFigure 12. Example 2 of Disagreement Between Specialists for Qualitative Evaluation

eFigure 13. Stacked Bar Chart Showing Distribution of the Stack Rank Positions for the Expert Gradings and the Model Segmentations

eFigure 14. Bland-Altman Plots Comparing Volumes of Individual Features Segmented Between Graders

eFigure 15. Matrix Showing the Number of Scans Segmented by the Model and Segmented by Neither, One or Both Human Expert Graders, per Feature

eFigure 16. Relationship Between Intergrader DSC and the Model-Grader DSC per Feature

eFigure 17. Distribution of Dice Similarity Coefficients (DSC) Stratified by 4 Ascending Mean Grader Volume Buckets

eFigure 18. Distribution of Dice Similarity Coefficients (DSC) for All Scans and Stratified by Scan Subgroup

eTable 1. Grader Experience and Set Assignment

eTable 2. Qualitative Evaluation Retinal Specialist Qualifications

eTable 3. Stack Rank Comparison Between the Model and Human Expert Graders Taking Into Account the Margins of Differences Within Each Stack

eTable 4. Magnitude of Difference Between Model and Next Expert Grader Where the Model Ranked Highest

eTable 5. Distribution of Specialist Likert Ratings for Model Segmentations

eTable 6. Distribution of Specialist Likert Ratings for Expert Gradings

eTable 7. Pairwise Intraclass Correlation Coefficient (ICC) With 95% CIs

eTable 8. Pairwise Intraclass Correlation Coefficient (ICC) and 95% CIs per Subgroup

eTable 9. Median and Interquartile Range of DSC for Scans Where the Feature Was Considered Present Across All 3 Segmentations

eTable 10. Median and Interquartile Range of DSC for Scans Where the Feature Was Considered Present Across All 3 Segmentations

This supplementary material has been provided by the authors to give readers additional information about their work.

eMethods

Grading protocol

Intraretinal fluid

IRF was defined as fluid cysts within the neurosensory retina. Cysts were visualized as intraretinal spaces with no reflectivity typically round/oval in shape, and usually located in the outer plexiform/nuclear retinal layers. IRF could present as one hypo-reflective space only or as multiple hypo-reflective spaces which were typically separated from each other by reflective septa. Lesions which were clearly cystoid spaces but contained some diffuse low to medium hyperreflective material (i.e., blood filled cysts) were delineated as IRF. IRF was graded if it was considered present by the grader with greater than 50% confidence at max 1.5x magnification. Small pockets of fluid away from the diseased areas were interpreted with caution. Diffuse retinal thickening without clearly distinguishable cystic spaces was not graded.

Subretinal fluid

SRF was defined as areas of non-reflectivity or moderate reflectivity between the posterior boundary of the neurosensory retina and the retinal pigment epithelium/Bruch's complex. The non-reflective space was typically semi-circular with tapered lateral extensions. Protruding photoreceptors were included within SRF. If SRF was labeled as present, continuous areas of SRF were graded irrespective of the variability in reflectivity of SRF. In cases of mixed features of SRF and hyper-reflective material in the subretinal space, a distinction was made between areas of SRF and areas of SHRM without overlap between the two.

Subretinal hyperreflective material (Set 2: AMD cases only)

SHRM was defined as an area of varied degrees of hyper reflectivity in the subretinal space. SHRM included both well-defined hyper-reflective material with clearly distinguishable boundaries from surrounding tissue (corresponding to fibrotic changes or vitelliform material) and ill-defined material that could not be distinguished from the surrounding tissue for at least 50% of its boundary outline (corresponding to exudation, blood, fibrinous material, elements of the choroidal neovascular membrane). The subretinal space was assessed along a spectrum as this can contain a mixture of fluid and SHRM. If the subretinal space was mainly hyporefective with low to medium hyperreflective material diffusely scattered within it (or above it), this was graded as SRF. SHRM was graded if the material was more concentrated/hyperreflective and/or clearly delineated. Caution was taken to not over-call moderately reflective material located above SRF, since it is more likely to correspond to thickened photoreceptors.

Pigment epithelial detachment (AMD cases only)

PED was defined as a separation between the retinal pigment epithelium layer and Bruch's membrane. If present, areas of fibrovascular PED, serous PED and drusen were segmented, regardless of size cut-offs. Fibrovascular PED was determined as a well-defined irregular elevation of the RPE with a deeper area of mild backscattering corresponding to the fibrous proliferation, topographically corresponding to areas of disease

activity or previous disease activity in treated cases. Serous PED was defined as an area of sharply demarcated, dome-shaped elevation of the RPE, with an area of homogenous hyporeflectivity in the sub-RPE space corresponding to an accumulation of fluid. Small and intermediate drusen were determined as discrete areas of RPE elevation with variable reflectivity. Larger, or confluent drusen were determined to be of hypo- to medium- reflectivity, and often visible as several connected dome-shaped elevations. Fibrosis was captured as either SHRM or PED, permitting the ambiguity in identifying the RPE. For set 1, each PED was segmented by type, i.e. fibrovascular, serous, drusen, and was collapsed into a singular PED for analysis. For set 2, all types of PED were segmented as a singular PED.

Disease severity grading definition

Wet age-related macular degeneration (AMD)

Mild disease: Absence of or presence of minimal IRF and/or SRF (few scattered intraretinal cysts and/or a thin layer of SRF), in the context of a fibrovascular PED

Moderate disease: Presence of IRF and/or SRF exceeding the threshold for mild disease, in the context of a fibrovascular PED with or without SHRM.

Severe disease: IRF and/or SRF exceeding the threshold for moderate disease (presence of large confluent cystic spaces and/or a high SRF volume visible as a large separation between the neurosensory retina and RPE), in the context of a fibrovascular PED with or without SHRM

Diabetic macular edema (DME)

Mild disease: Absence of or presence of minimal IRF and/or SRF (few scattered intraretinal cysts) and/or a thin layer of SRF.




Moderate disease: Presence of IRF (several cystic spaces in clusters or in a diffuse pattern) exceeding the threshold for mild disease, with or without SRF.

Severe disease: Presence of IRF (large confluent cystic spaces) exceeding the threshold for moderate disease, with or without SRF.

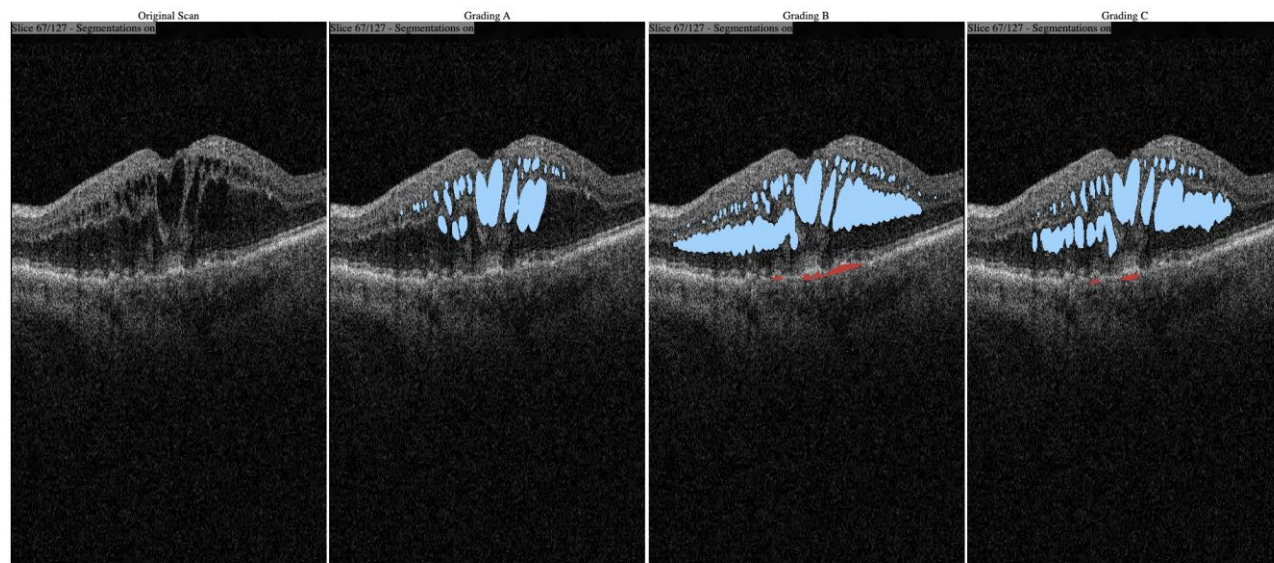
Supplementary Figures

Qualitative evaluation viewer

Features

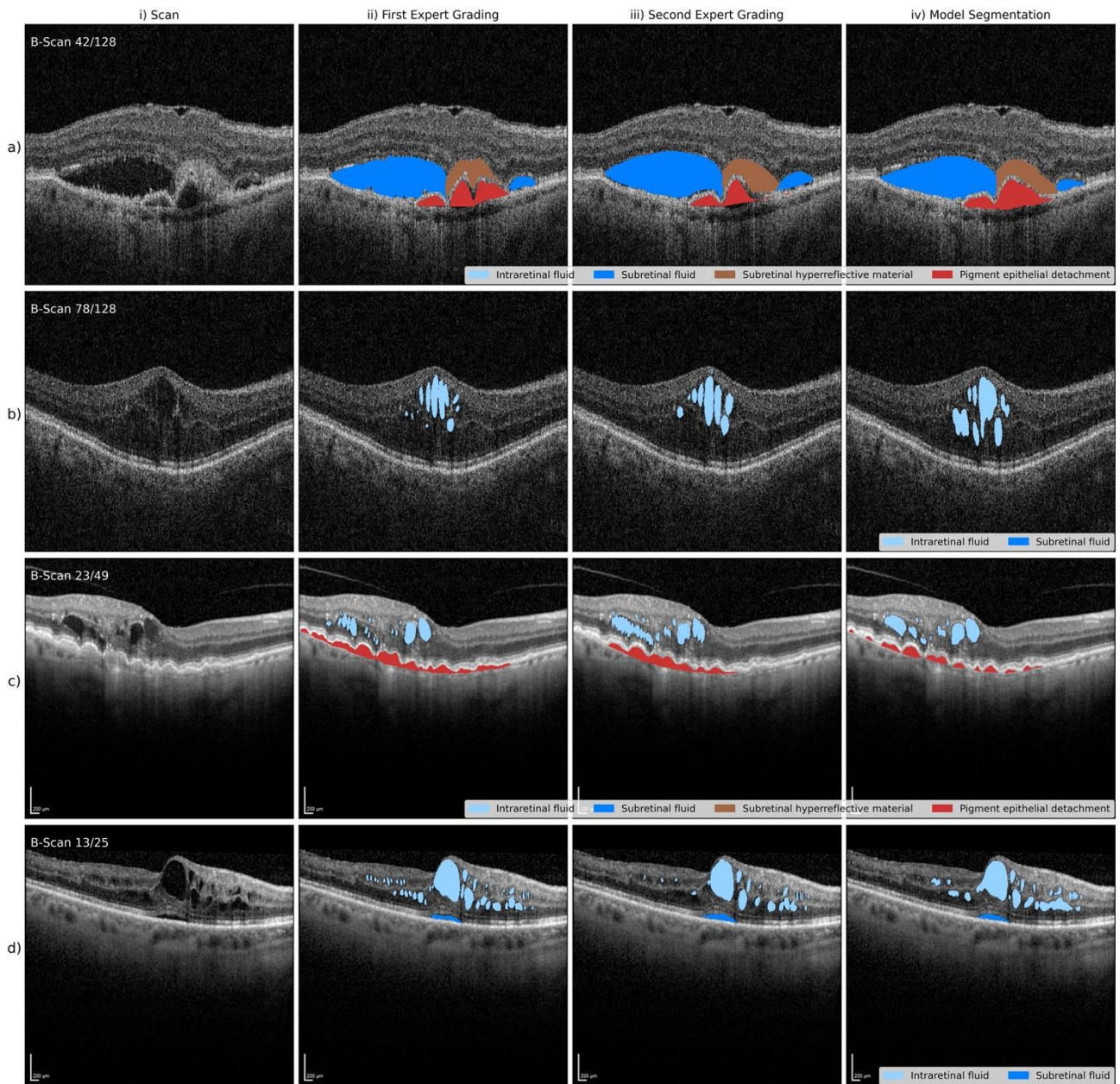
- Intraretinal Fluid 
- Subretinal Fluid 
- PED 
- Subretinal Hyper Reflective Material Not graded

Gradings



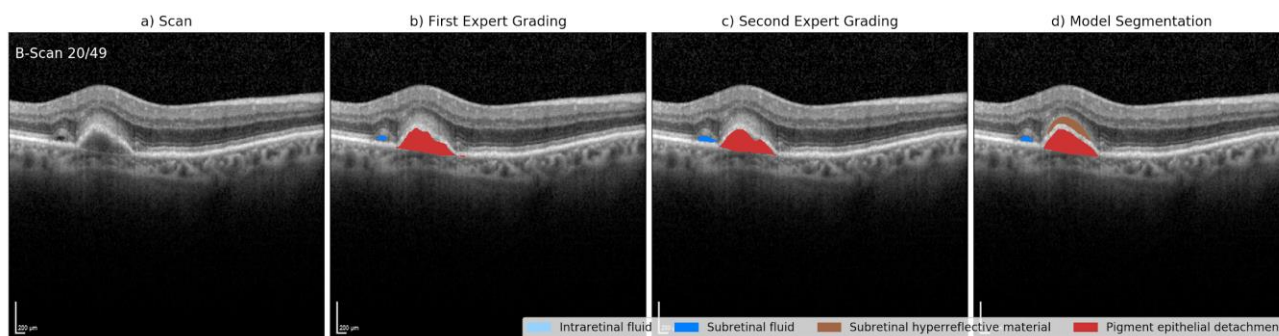
Controls:
Scroll in scan = Move through slices
Right/Left = Next/Previous slice
S = Toggle segmentations
F = Jump to first slice
L = Jump to last slice
C = Jump to central slice

eFigure 1. Custom viewer used to assess the 3 segmentations for the qualitative evaluation. Gradings A, B, and C represented expert and model segmentations in a randomised order. In this particular case, Grading A and Grading B were manual gradings, and Grading C was the model prediction.

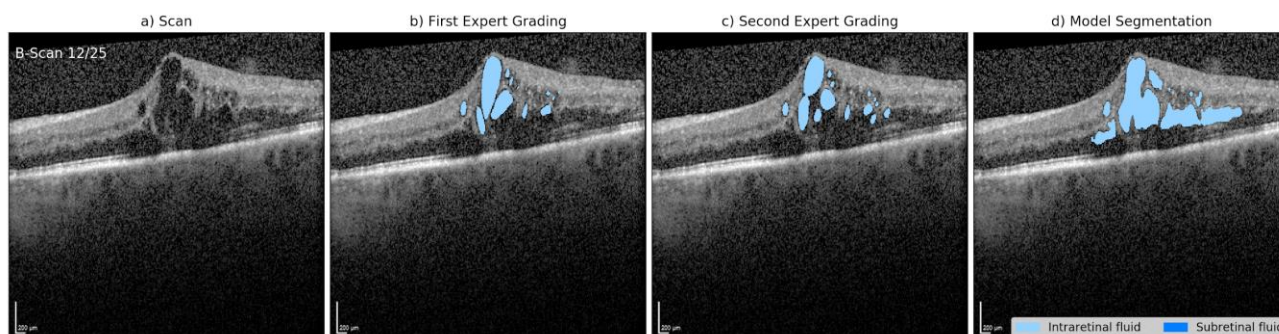


eFigure 2. Example segmentations from: a) Set 2: Topcon-AMD, b) Set 2: Topcon-DME, c) Set 2: Heidelberg-AMD, and d) Set 2: Heidelberg-DME with: i) The OCT B-scan, ii, iii) the two expert gradings, and iv) the model segmentation. Up to 2 features were segmented for DME scans: intraretinal fluid and subretinal fluid. For AMD scans, up to 4 features were segmented: intraretinal fluid, subretinal fluid, subretinal hyperreflective material, and pigment epithelial detachment. See the Video for whole volume segmentations.

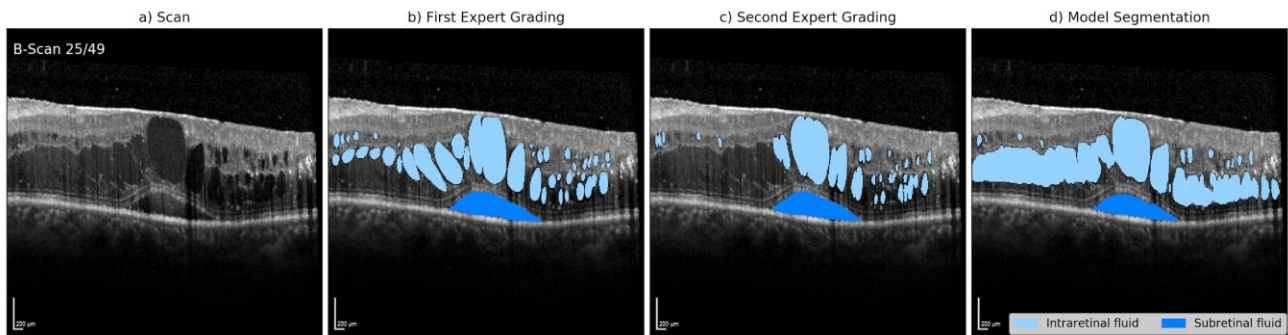
Model success cases



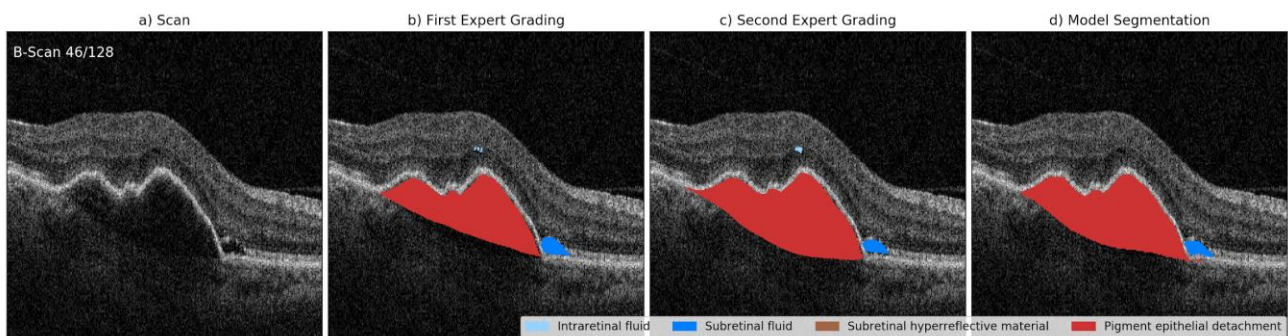
eFigure 3. Example of a scan where the model was qualitatively rated higher than the two manual segmentations. a) Raw OCT of an individual with wet AMD, taken on a Heidelberg device. b) and c) are expert gradings, and d) is the model grading. The model segments fluid and PED in a similar pattern to the expert gradings. Intergrader DSC for IRF and PED is 0.65 and 0.79, respectively. Model-grader DSC was 0.67 and 0.79 (for both grader 1 and 2) for IRF and PED, respectively. Both expert gradings did not segment any SHRM. However, all specialists felt that the model segmentation was most representative. See the Video for a video representation of this case.



eFigure 4. Example of a scan where the model was qualitatively rated higher than the two manual segmentations. a) Raw OCT of an individual with DME, taken on a Heidelberg device. The quality of the scan is hampered by speckle noise which makes it challenging to judge intraretinal cyst boundaries; b) and c) are expert gradings, and d) is the model grading. For this scan, IRF was the only feature segmented in all three gradings. The model segments more IRF (0.64mm^3) than the expert gradings (0.15mm^3 and 0.13mm^3). The intergrader DSC is 0.76, while the model-grader DSC is 0.37 and 0.32. The specialists ranked the model segmentations as first, and considerably better than the manual gradings. See the Video for a video representation of this case.

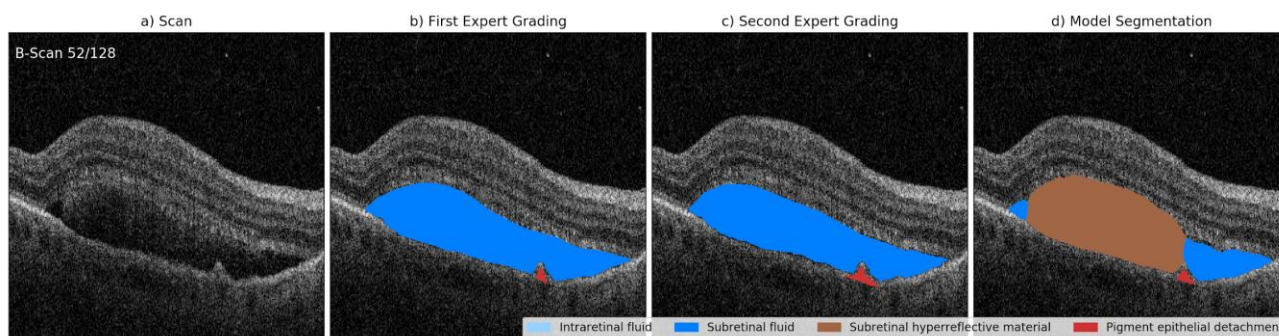


eFigure 5. Example of a scan where the model was ranked as most representative. a) Raw OCT of an individual with severe DME, taken on a Heidelberg device. The scan shows significant IRF and a cuff of SRF; b) and c) are expert gradings, both segmenting similar amounts of SRF, and variable amounts of IRF; d) the model correctly segments a greater volume of IRF particularly in the parafoveal and perifoveal areas. In severe DME, these regions can be challenging to ambiguate from retinal thickening. The specialists ranked the model grading as most representative, and rated it higher than the manual gradings. See the Video for a video representation of this case.

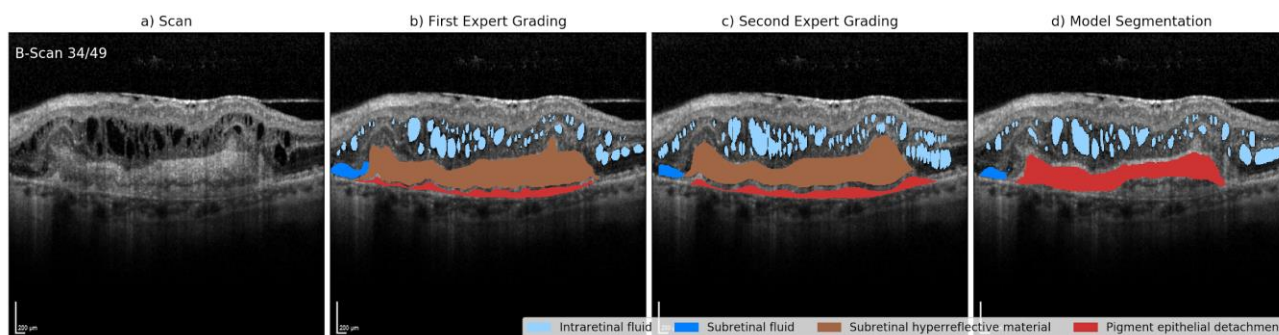


eFigure 6. Example of a scan where all three gradings were rated highly. a) Raw OCT of an individual with wet AMD, taken on a Topcon device; b) and c) are expert gradings, both segmenting a large PED, an area of SRF, and minimal IRF (segmented only in this B-scan); d) the model segments PED and SRF to a similar extent, but does not segment any IRF. The intergrader DSC is 0.89, and 0.89 for SRF, and PED, respectively. Model-grader DSC was 0.90 for SRF (for both expert graders), and 0.90–0.92 for PED. The specialists agreed or strongly agreed that they would be satisfied to use each of these segmentations within clinical practice. The model was ranked second by Specialist 2 and 3, and third by Specialist 1. See the Video for a video representation of this case.

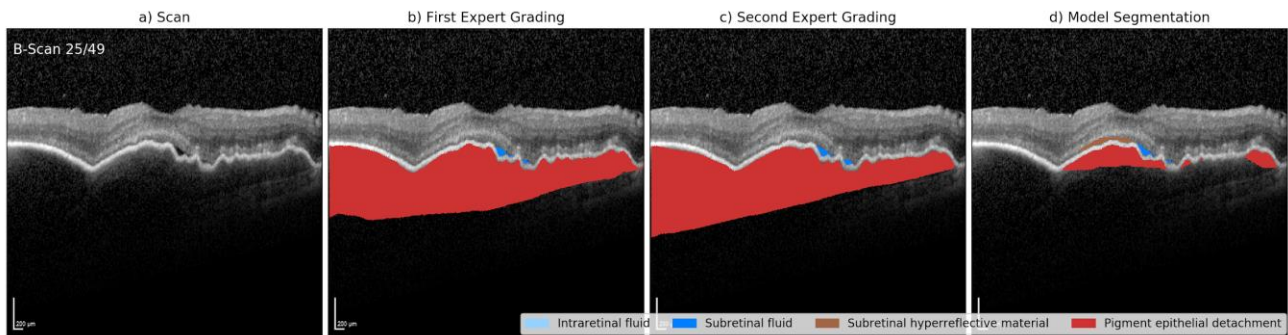
Challenging cases and model failures



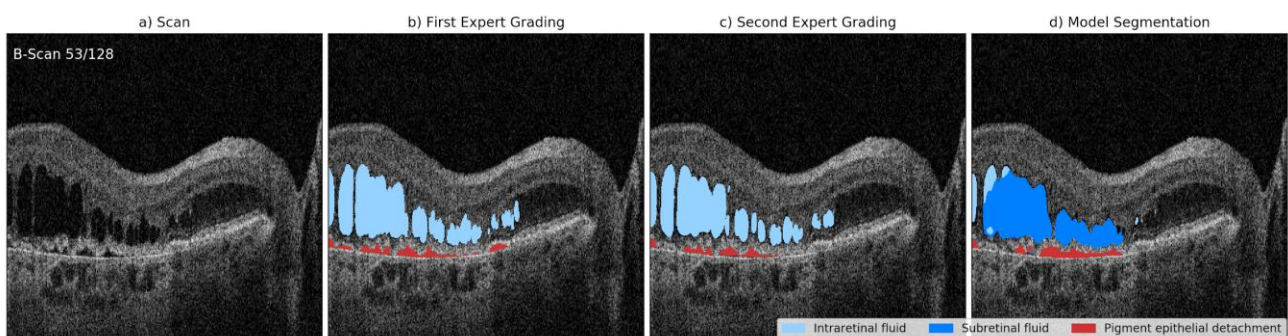
eFigure 7. Example of a challenging case resulting in large disagreement in SRF volume between model and manual gradings. a) Raw OCT from set 1 of an individual with wet AMD, taken on a Topcon device. The scan shows a neurosensory detachment with two distinct morphological compartments comprising a large area of subretinal haemorrhage showing some scattered reflectivity, with surrounding SRF which has greater hyporeflectivity; b) and c) are manual gradings, both segmenting a large area of SRF. For this set, graders were not asked to segment SHRM; d) model segmentation clearly recognises the morphological compartments but segments the haemorrhage as SHRM. This may be justifiable considering that the model has not been trained to classify haemorrhage. The specialists gave the model segmentation the lowest rating. See the Video for a video representation of this case.



eFigure 8. Example of large disagreement in SHRM and PED volume between model and manual gradings. a) Raw OCT of an individual with wet AMD, taken on a Heidelberg device. The scan shows an area of fibrosis and disorganised tissue. This area likely represents a mixture of PED and SHRM with undefined boundaries. b) and c) are manual gradings, both segmenting SHRM and PED, using a region of hyporeflectivity as a boundary between the two features; d) the model segments this region as predominantly PED, using the same region of hyporeflectivity as a boundary for the outer extent of the PED. The specialists ranked the model third and rated it 1 or 2 on the Likert scale.

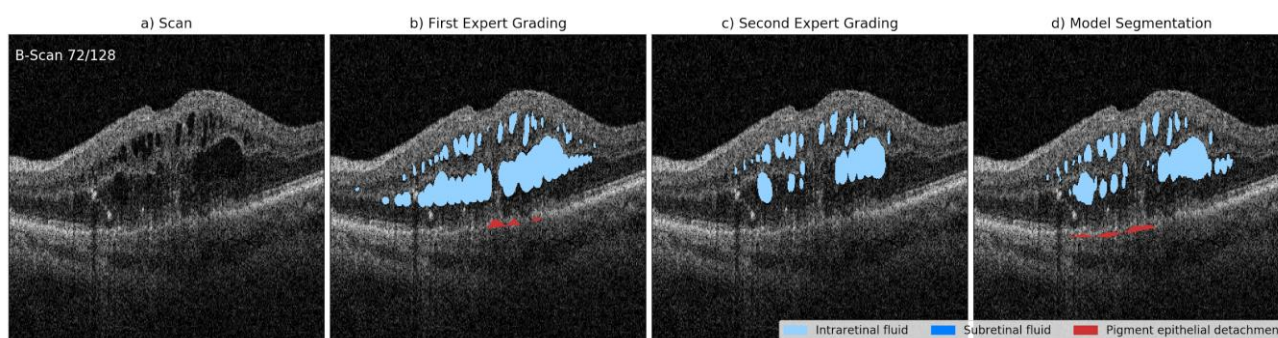


eFigure 9. Example of model failure and largest disagreement in PED volume between the model and expert graders. a) A large PED is visible on the raw OCT of an individual with wet AMD, taken on a Heidelberg device. Bruch's membrane is undefined except for the far right where the PED is slightly shallower. This makes it difficult to infer the outer boundary of the PED; b) and c) are expert gradings, both estimating large volumes of PED with disagreement on the extent of the outer boundary; d) the model segments a very shallow PED up to the point of OCT penetration, under calling the volume and extent of PED. The specialists ranked the model third and rated it 1 or 2 on the Likert scale. See the Video for a video representation of this case.

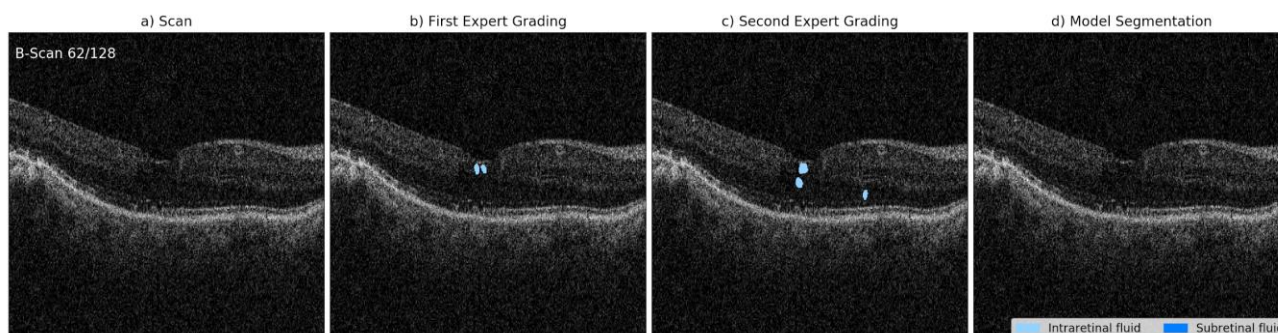


eFigure 10. Example of a model failure and large disagreement in IRF/SRF volume. a) Raw OCT of severe wet AMD from Set 1: Topcon-AMD; b) and c) are expert gradings - both segment a large volume of IRF; d) the model incorrectly segments this fluid compartment as SRF as the retinal tissue above the RPE is indistinct. The specialists ranked the model segmentation as third. Specialist 3 disagreed that they would be satisfied to use any of the 3 presented segmentations in clinical practice. Specialist 1 and 2 were either neutral or agreed with the expert gradings and disagreed with the model segmentation. See the Video for a video representation of this case.

Disagreements between specialists

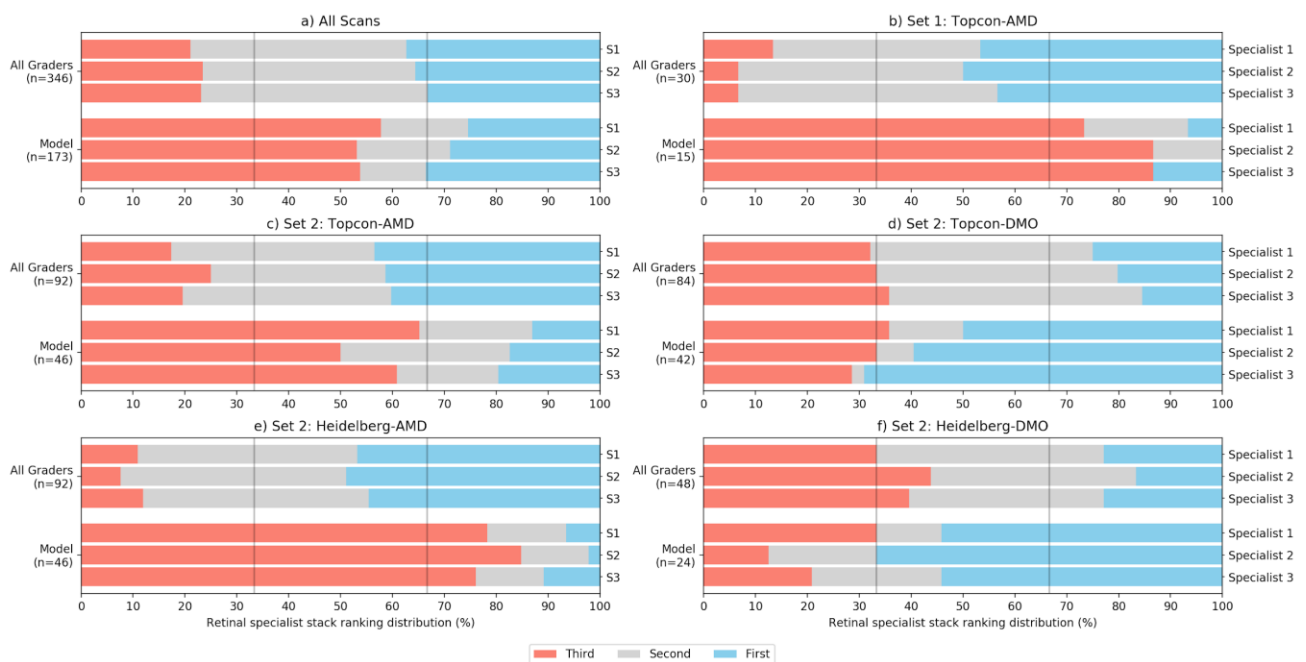


eFigure 11. Example of a disagreement between specialists for qualitative evaluation. a) Raw OCT from Set 1: Topcon-AMD shows severe wet AMD with significant intraretinal fluid with shallow PED in this B-scan; b) and c) are expert gradings and d) is the model output. The integrader and model-grader DSC was similar for IRF and PED. Specialists 1 and 2 either agreed or strongly agreed that they would be satisfied to use the 3 segmentations in their clinical practice. Specialist 3 agreed that the first expert grading was satisfactory, but strongly disagreed with the second expert grading and model grading. See the Video for a video representation of this case.



eFigure 12. Example of disagreement between specialists for qualitative evaluation. a) Raw OCT from Set 2: Topcon-DME of a mild DME case; b) and c) are expert gradings, both segmenting few intraretinal cysts with little overlap; d) the model did not segment any intraretinal fluid. Specialist 1 disagreed that the model output was satisfactory, yet Specialists 2 and 3 agreed that it was. Specialists 1 and 2 both ranked the model third, and Specialist 3 ranked the model second. See the Video for a video representation of this case.

Qualitative evaluation: stack rank



eFigure 13. Stacked bar chart showing distribution of the stack rank positions for the expert gradings and the model segmentations. Each bar represents a retinal specialist (S1-S3), for a) all scans collectively and b-f) each set of scans. This takes into account even slight differences within each stack rank. Lines are overlaid at 33.33% and 66.67% to indicate the distribution of chance.

Quantitative comparison of volume segmented between expert graders

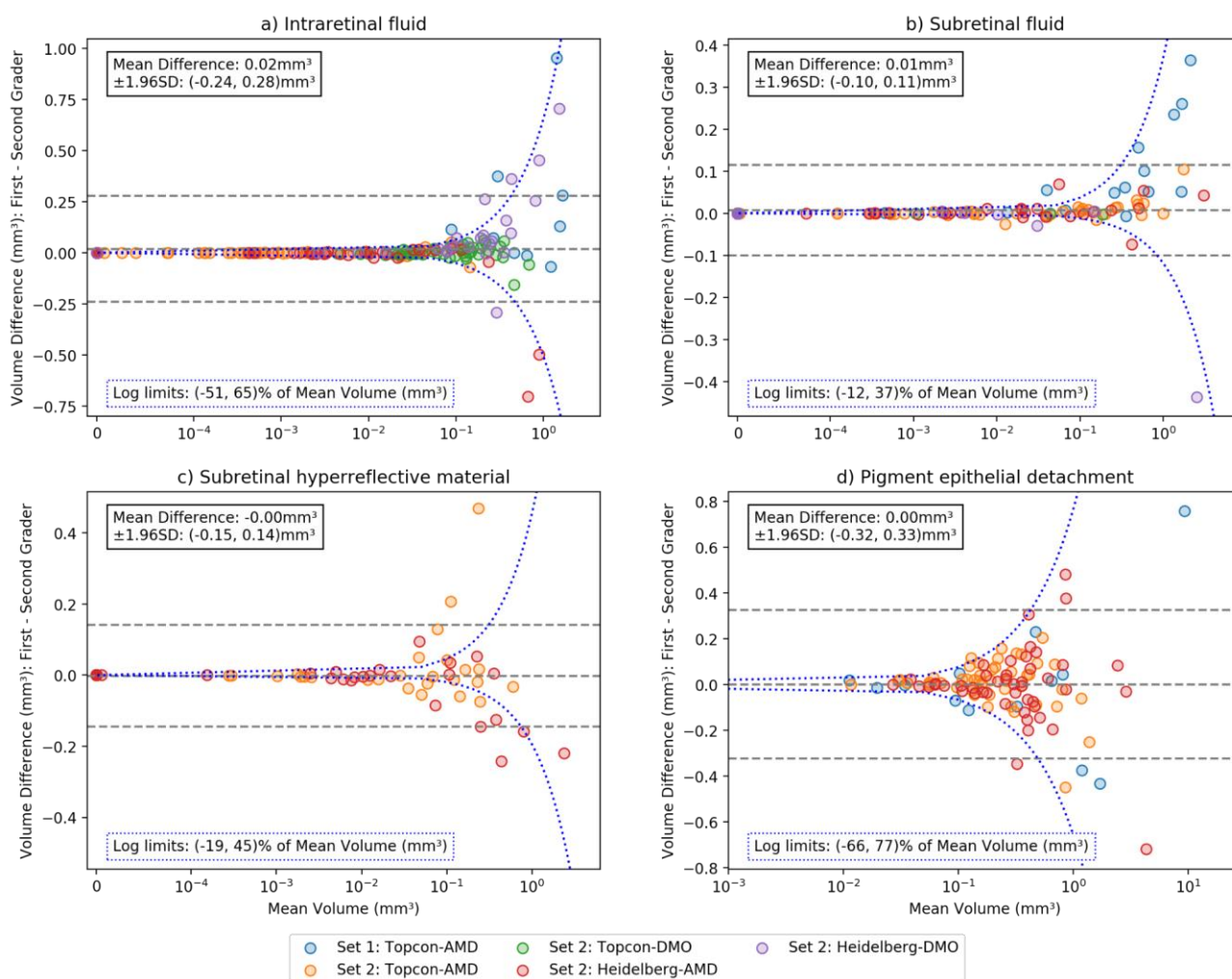
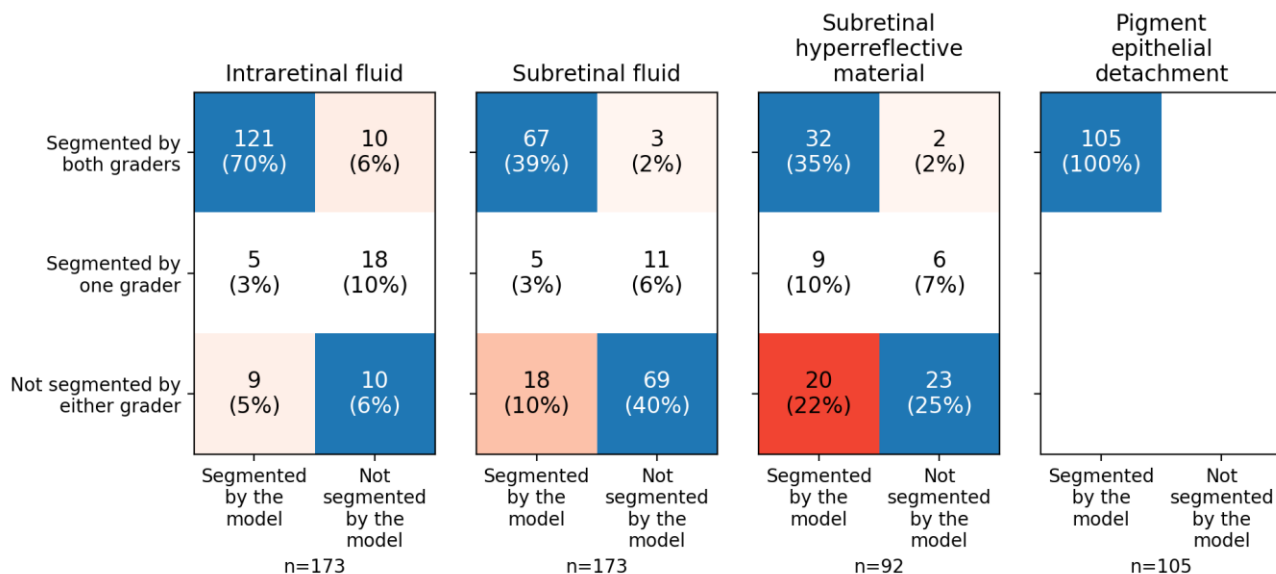


Figure 14. Bland-Altman plots comparing volumes of individual features segmented between graders. The mean value of the difference and the 95% limits of agreement (mean difference ± 1.96 SD of the difference) are plotted with black dashed lines. Given the differences are related to magnitude of the mean volume, the limits of agreement are also calculated after log-transforming the data, and are plotted in the linear space, as a ratio of the mean volume, with blue dotted lines.

Comparison of features segmented between graders and model



eFigure 15. Matrix showing the number of scans (%) segmented by the model and segmented by neither, one or both human expert graders, per feature. Only scans in the top left where both human expert graders and the model segmented at least one voxel in the volume are used for Dice Similarity Coefficient analysis.

Intergrader versus model-grader Dice Similarity Coefficient

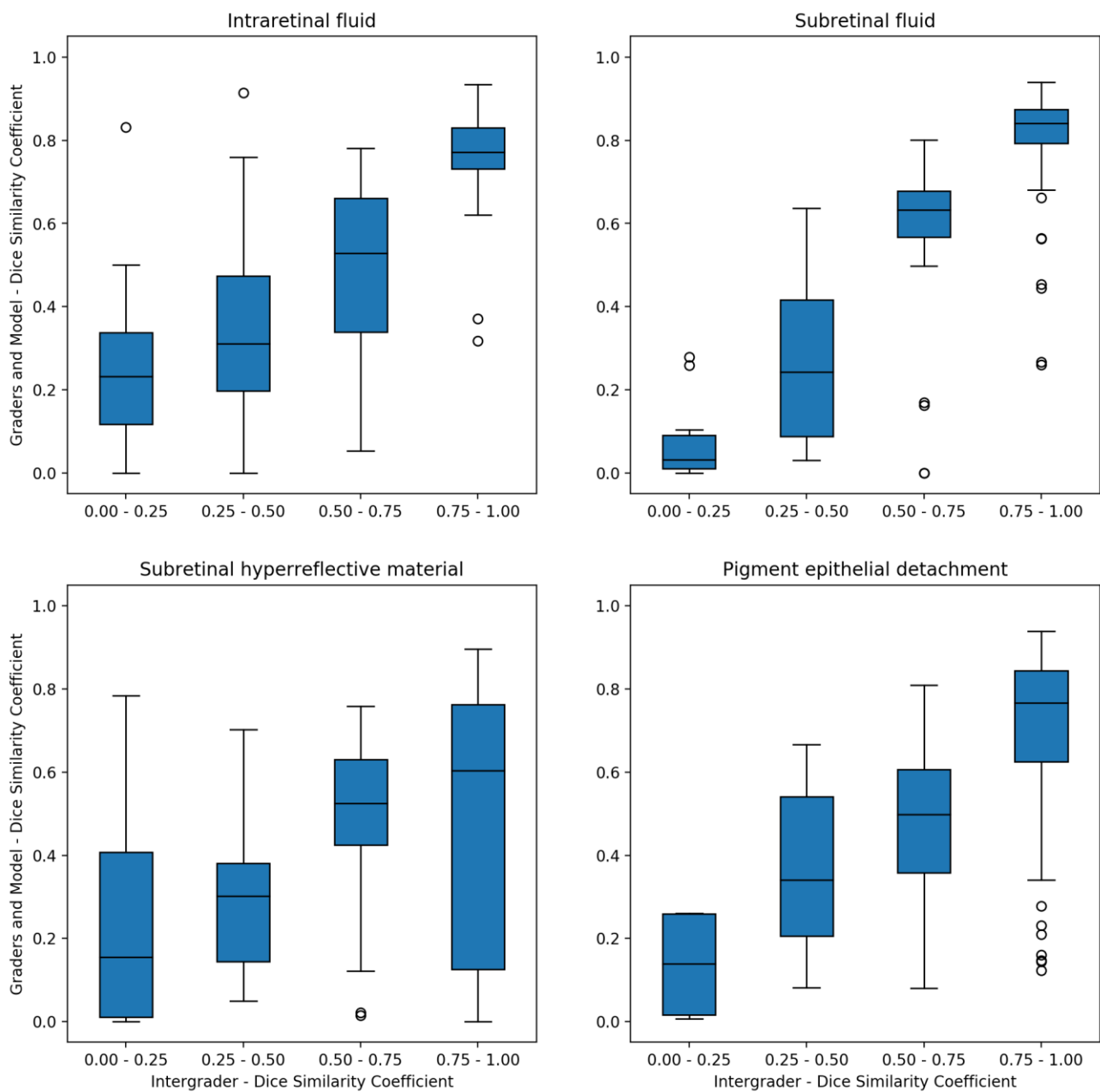


Figure 16. Relationship between intergrader DSC and the model-grader DSC per feature. Generally, as intergrader DSC increases, median model-grader DSC also increases.

Dice Similarity Coefficients boxplots stratified by volume

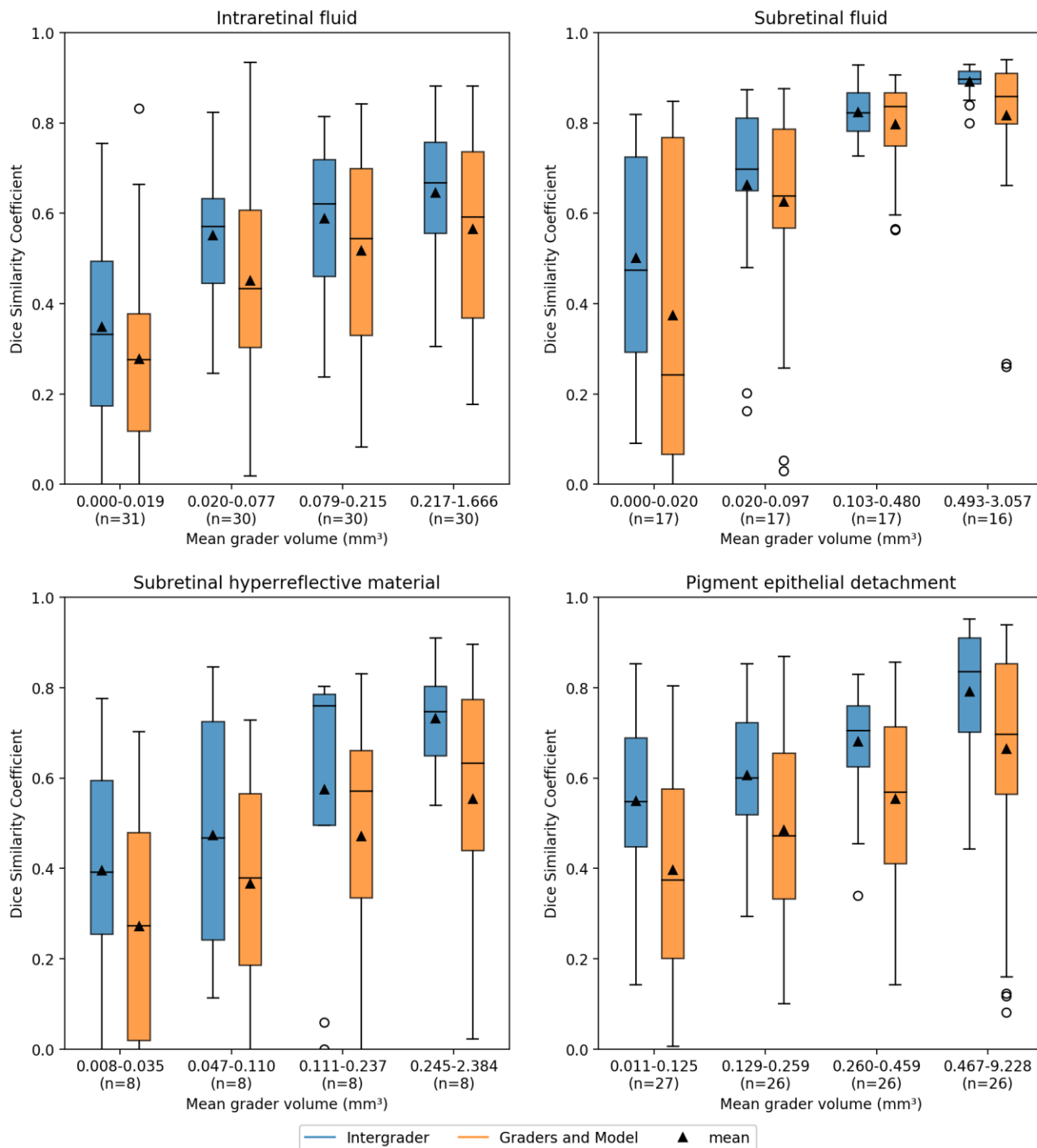
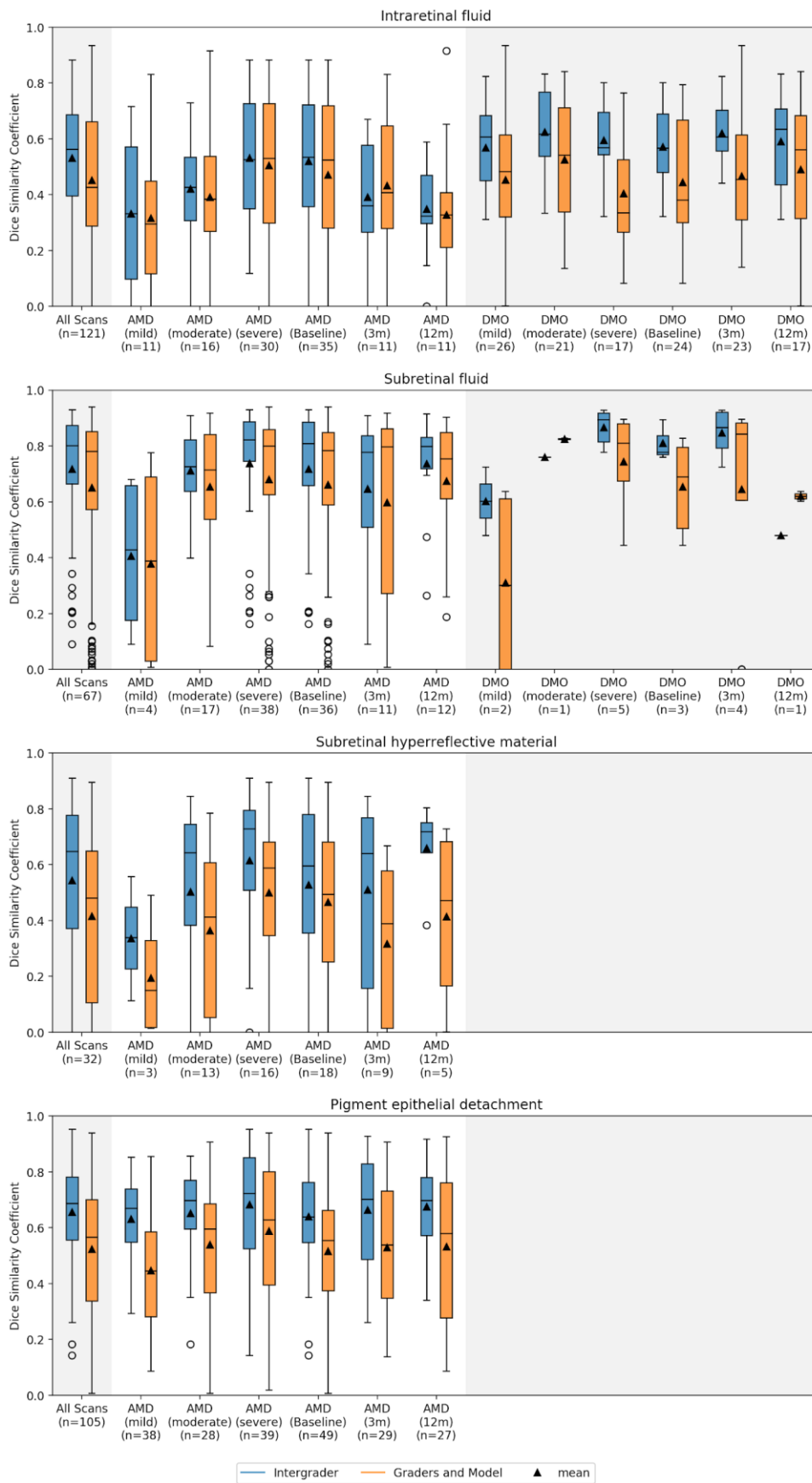


Figure 17. Distribution of Dice Similarity Coefficients (DSC) stratified by four ascending mean grader volume buckets, for: a) intraretinal fluid; b) subretinal fluid; c) subretinal hyperreflective material; and d) pigment epithelial detachment. Boxes display the median and interquartile range. The whiskers extend up to 1.5*interquartile range beyond the upper and lower quartiles; the isolated circles fall outside of this range. The black triangles represent the mean DSC.

Dice Similarity Coefficients by subgroup



eFigure 18. Distribution of Dice Similarity Coefficients (DSC) for all scans and stratified by scan subgroup for: a) intraretinal fluid; b) subretinal fluid; c) subretinal hyperreflective material; and d) pigment epithelial detachment. Boxes display the median and interquartile range. The whiskers indicate most extreme, non-outlier data points. Where data lies outside 1.5 x interquartile range it is represented as a circular flier. The black triangles represent the mean DSC.

Supplementary Tables

Graders and retinal specialists

Grader	Qualification	Grading experience	Sets graded
A	Qualified Optometrist	2 years	Set 1: Topcon-AMD
B	Qualified Optometrist	4 years	Set 1: Topcon-AMD
C	Science PhD (non-medical degree)	6 years	Set 2: Topcon-DME Set 2: Heidelberg-AMD Set 2: Heidelberg-DME
D	Medical degree	8 years	Set 2: Topcon-AMD Set 2: Heidelberg-AMD Set 2: Heidelberg-DME
E	Qualified Optometrist	3 years	Set 2: Topcon-AMD
F	Qualified Optometrist	2 years	Set 2: Topcon-DME

eTable 1. Grader experience and set assignment

Retinal specialist	Qualification
1	Ophthalmologist with Medical Retina fellowship training
2	Consultant Ophthalmologist in Medical Retina
3	Ophthalmologist with Medical Retina fellowship training

eTable 2. Qualitative evaluation retinal specialist qualifications

Qualitative evaluation: stack rank

Dataset	Number of Scans	Differences Accepted as Equivalent	Model segmentations compared to expert gradings			
			Better n (%)	Comparable n (%)	Worse n (%)	Better or comparable to at least one expert grading n (%; 95% CI)
All scans	173	None	49 (28%)	29 (17%)	95 (55%)	78 (45%; 38-53%)
All scans	173	Slight	40 (23%)	81 (47%)	52 (30%)	127 (73%; 66-79%)
All scans	173	Slight and Moderate	15 (9%)	133 (77%)	25 (14%)	149 (86%; 80-90%)
Set 1: Topcon-AMD	15	None	1 (7%)	1 (7%)	13 (87%)	2 (13%; 4-38%)
Set 1: Topcon-AMD	15	Slight	1 (7%)	7 (47%)	7 (47%)	9 (60%; 36-80%)
Set 1: Topcon-AMD	15	Slight and Moderate	0 (0%)	9 (60%)	6 (40%)	9 (60%; 36-80%)
Set 2: Topcon-AMD	46	None	5 (11%)	14 (30%)	27 (59%)	19 (41%; 28-56%)
Set 2: Topcon-AMD	46	Slight	6 (13%)	27 (59%)	13 (28%)	35 (76%; 62-86%)
Set 2: Topcon-AMD	46	Slight and Moderate	3 (7%)	38 (83%)	5 (11%)	41 (89%; 77-95%)
Set 2: Topcon-DME	42	None	26 (62%)	3 (7%)	13 (31%)	29 (69%; 54-81%)
Set 2: Topcon-DME	42	Slight	15 (36%)	21 (50%)	6 (14%)	36 (86%; 72-93%)
Set 2: Topcon-DME	42	Slight and Moderate	7 (17%)	33 (79%)	2 (5%)	40 (95%; 84-99%)
Set 2: Heidelberg-AMD	46	None	2 (4%)	6 (13%)	38 (83%)	8 (17%; 9-31%)
Set 2: Heidelberg-AMD	46	Slight	5 (11%)	16 (35%)	25 (54%)	24 (52%; 38-66%)
Set 2: Heidelberg-AMD	46	Slight and Moderate	1 (2%)	33 (72%)	12 (26%)	35 (76%; 62-86%)
Set 2: Heidelberg-DME	24	None	15 (63%)	5 (21%)	4 (17%)	20 (83%; 64-93%)
Set 2: Heidelberg-DME	24	Slight	13 (54%)	10 (42%)	1 (4%)	23 (96%; 80-99%)
Set 2: Heidelberg-DME	24	Slight and Moderate	4 (17%)	20 (83%)	0 (0%)	24 (100%; 86-100%)

Table 3. Stack rank comparison between the model and human expert graders taking into account the margins of differences within each stack. The model is considered better if: the majority of specialists ranked it higher than the two expert gradings; or higher than one and the same as one expert grading. The model is considered worse if: the majority of specialists ranked it lower than the two expert gradings; or lower than one and the same as one expert grading. The model is considered comparable if the majority of specialists neither ranked it better or worse. The model was considered better or comparable to at least one expert grading if the majority of specialists ranked it higher than or the same as one of the expert gradings.

	Number of scans where model ranked highest (total 173)	Slight difference to next expert, n (%)	Moderate difference to next expert, n (%)	Considerable difference to next expert, n (%)
Specialist 1	44	33 (75)	9 (20)	2 (5)
Specialist 2	50	22 (44)	15 (30)	13 (26)
Specialist 3	58	23 (40)	15 (26)	20 (23)

eTable 4. Magnitude of difference between model and next expert grader where the model ranked highest.

Qualitative evaluation: Likert ratings

Dataset	No. of Scans	Strongly disagree n (%)	Disagree n (%)	Neither agree or disagree n (%)	Agree n (%)	Strongly agree n (%)	Positive n (%; 95% CI)	Neutral or Positive n (%; 95% CI)
All scans	173	8 (5%)	30 (17%)	50 (29%)	83 (48%)	2 (1%)	85 (49%; 42-57%)	135 (78%; 71-84%)
Set 1: Topcon-AMD	15	3 (20%)	4 (27%)	5 (33%)	3 (20%)	0 (0%)	3 (20%; 7-45%)	8 (53%; 30-75%)
Set 2: Topcon-AMD	46	2 (4%)	4 (9%)	19 (41%)	20 (43%)	1 (2%)	21 (46%; 32-60%)	40 (87%; 74-94%)
Set 2: Topcon-DME	42	0 (0%)	6 (14%)	8 (19%)	27 (64%)	1 (2%)	28 (67%; 52-79%)	36 (86%; 72-93%)
Set 2: Heidelberg-AMD	46	3 (7%)	14 (30%)	15 (33%)	14 (30%)	0 (0%)	14 (30%; 19-45%)	29 (63%; 49-75%)
Set 2: Heidelberg-DME	24	0 (0%)	2 (8%)	3 (13%)	19 (79%)	0 (0%)	19 (79%; 60-91%)	22 (92%; 74-98%)

eTable 5. Distribution of specialist Likert ratings for model segmentations. The number and percentage of scans that a majority of specialists gave a positive (rated 4 or 5) or 'neutral or positive' (rated 3, 4 or 5) rating is shown.

Dataset	Number of Gratings	Strongly disagree n (%)	Disagree n (%)	Neither agree or disagree n (%)	Agree n (%)	Strongly agree n (%)	Positive n (%; 95% CI)	Neutral or Positive n (%; 95% CI)
All scans	346	0 (0%)	37 (11%)	84 (24%)	214 (62%)	11 (3%)	225 (65%; 60-70%)	309 (89%; 86-92%)
Set 1: Topcon-AMD	30	0 (0%)	2 (7%)	7 (23%)	20 (67%)	1 (3%)	21 (70%; 52-83%)	28 (93%; 79-98%)
Set 2: Topcon-AMD	92	0 (0%)	7 (8%)	12 (13%)	70 (76%)	3 (3%)	73 (79%; 70-86%)	85 (92%; 85-96%)
Set 2: Topcon-DME	84	0 (0%)	18 (21%)	31 (37%)	34 (40%)	1 (1%)	35 (42%; 32-52%)	66 (79%; 69-86%)
Set 2: Heidelberg-AMD	92	0 (0%)	2 (2%)	13 (14%)	71 (77%)	6 (7%)	77 (84%; 75-90%)	90 (98%; 92-99%)
Set 2: Heidelberg-DME	48	0 (0%)	8 (17%)	21 (44%)	19 (40%)	0 (0%)	19 (40%; 27-54%)	40 (83%; 70-91%)

eTable 6. Distribution of specialist Likert ratings for expert gradings. Each scan is represented twice, once per expert grading. The number and percentage of scans that a majority of specialists gave a positive (rated 4 or 5) or 'neutral or positive' (rated 3, 4 or 5) rating is shown.

Intraclass correlation coefficients by subgroup

	Intraretinal fluid		Subretinal fluid		Subretinal hyperreflective material		Pigment epithelial detachment	
	Intergrader	Model-Grader	Intergrader	Model-Grader	Intergrader	Model-Grader	Intergrader	Model-Grader
All Scans	0.90 (0.83-0.96)	0.46 (0.35-0.62)	0.99 (0.99-1.00)	0.96 (0.90-0.99)	0.97 (0.72-0.99)	0.33 (0.08-0.96)	0.99 (0.94-0.99)	0.89 (0.35-0.99)
Set 1: Topcon-AMD	0.90 (0.74-0.99)	0.93 (0.85-0.99)	0.98 (0.97-0.99)	0.96 (0.89-0.99)	N/A	N/A	0.99 (0.90-1.00)	1.00 (0.84-1.00)
Set 2: Topcon-AMD	0.89 (0.82-0.96)	0.85 (0.57-0.97)	1.00 (1.00-1.00)	0.98 (0.84-1.00)	0.76 (0.26-0.98)	0.82 (0.44-0.96)	0.94 (0.86-0.97)	0.89 (0.71-0.96)
Set 2: Topcon-DME	0.97 (0.94-0.99)	0.33 (0.22-0.63)	1.00 (1.00-1.00)	0.92 (0.00-0.99)	N/A	N/A	N/A	N/A
Set 2: Heidelberg-AMD	0.74 (0.56-0.97)	0.73 (0.08-0.96)	1.00 (0.98-1.00)	0.99 (0.83-1.00)	0.99 (0.86-0.99)	0.24 (0.05-0.98)	0.98 (0.82-0.99)	0.46 (0.15-0.87)
Set 2: Heidelberg-DME	0.83 (0.65-0.88)	0.31 (0.17-0.48)	0.99 (0.59-1.00)	0.84 (0.10-0.97)	N/A	N/A	N/A	N/A

eTable 7. Pairwise intraclass correlation coefficient (ICC) with 95% confidence intervals. The ICC measures the agreement of volumes segmented, between graders (intergrader) and between the model and graders for each scan (model-grader).

	Intraretinal fluid		Subretinal fluid		Subretinal hyperreflective material		Pigment epithelial detachment	
	Intergrader	Model-Grader	Intergrader	Model-Grader	Intergrader	Model-Grader	Intergrader	Model-Grader
Topcon (all)	0.94 (0.83-0.99)	0.67 (0.44-0.88)	0.99 (0.99-1.00)	0.98 (0.94-0.99)	0.76 (0.26-0.98)	0.82 (0.44-0.96)	0.99 (0.91-1.00)	0.99 (0.84-1.00)
Heidelberg (all)	0.83 (0.71-0.89)	0.39 (0.27-0.53)	0.99 (0.98-1.00)	0.95 (0.81-1.00)	0.99 (0.86-0.99)	0.24 (0.05-0.98)	0.98 (0.82-0.99)	0.46 (0.15-0.87)
Heidelberg 25 B-scans	0.87 (0.68-0.95)	0.30 (0.21-0.44)	0.99 (0.98-1.00)	0.85 (0.81-0.99)	0.91 (0.62-0.98)	0.84 (0.56-0.95)	0.98 (0.78-1.00)	0.87 (0.44-0.91)
Heidelberg 49 B-scans	0.81 (0.62-0.88)	0.44 (0.29-0.90)	1.00 (0.95-1.00)	1.00 (0.78-1.00)	0.99 (0.86-1.00)	0.19 (0.02-1.00)	0.97 (0.71-0.99)	0.25 (0.09-0.71)
AMD (all)	0.91 (0.79-0.98)	0.92 (0.77-0.98)	0.99 (0.99-1.00)	0.98 (0.95-0.99)	0.97 (0.72-0.99)	0.33 (0.08-0.96)	0.99 (0.94-0.99)	0.89 (0.35-0.99)
AMD (mild)	0.95 (0.25-0.99)	0.28 (-0.05-0.74)	0.77 (0.22-0.90)	0.83 (0.05-0.96)	0.35 (0.15-0.91)	0.42 (-0.10-0.64)	0.81 (0.63-0.92)	0.47 (0.24-0.67)
AMD (moderate)	0.88 (0.73-0.96)	0.11 (0.00-0.92)	1.00 (0.99-1.00)	0.98 (0.93-1.00)	0.81 (0.41-0.93)	0.74 (0.44-0.96)	0.86 (0.75-0.91)	0.38 (0.13-0.62)
AMD (severe)	0.89 (0.75-0.98)	0.94 (0.87-0.98)	0.99 (0.98-1.00)	0.98 (0.94-0.99)	0.97 (0.65-0.99)	0.20 (-0.01-0.97)	0.99 (0.94-0.99)	0.89 (0.23-0.99)
AMD (baseline)	0.90 (0.75-0.98)	0.94 (0.88-0.98)	0.99 (0.98-1.00)	0.98 (0.96-0.99)	0.97 (0.63-0.99)	0.25 (0.02-0.97)	0.99 (0.89-0.99)	0.89 (0.19-0.99)
AMD (3m)	0.95 (0.65-0.97)	0.06 (0.00-0.97)	1.00 (0.99-1.00)	1.00 (0.96-1.00)	0.93 (0.46-0.99)	0.90 (0.60-0.94)	0.99 (0.92-1.00)	0.91 (0.71-0.94)
AMD (12m)	0.81 (0.48-0.90)	0.67 (-0.03-0.86)	1.00 (0.99-1.00)	0.62 (0.40-1.00)	0.93 (0.83-1.00)	0.68 (0.33-0.98)	0.92 (0.66-0.97)	0.82 (0.22-0.95)
DME (all)	0.87 (0.80-0.92)	0.34 (0.23-0.49)	0.99 (0.91-1.00)	0.84 (0.58-0.98)	N/A		N/A	
DME (baseline)	0.78 (0.58-0.89)	0.22 (0.10-0.33)	0.99 (0.44-1.00)	0.84 (-0.01-0.89)				
DME (3m)	0.89 (0.87-0.99)	0.43 (0.18-0.59)	1.00 (0.95-1.00)	0.99 (0.00-0.99)				
DME (12m)	0.93 (0.88-1.00)	0.53 (0.36-0.97)	0.62 (0.59-1.00)	0.82 (0.00-0.93)				
DME (mild)	0.96 (0.92-0.97)	0.59 (0.37-0.86)	0.62 (0.60-1.00)	0.83 (0.00-1.00)				
DME (moderate)	0.72 (0.30-0.94)	0.09 (-0.13-0.40)	0.61 (0.00-1.00)	0.03 (-0.01-0.95)				
DME (severe)	0.80 (0.60-0.86)	0.22 (0.06-0.36)	0.98 (0.98-1.00)	0.84 (0.58-0.98)				
Left	0.92 (0.81-1.00)	0.68 (0.37-0.94)	0.99 (0.98-1.00)	0.99 (0.97-1.00)	0.35 (-0.01-1.00)	0.69 (0.03-0.98)	0.99 (0.87-1.00)	1.00 (0.78-1.00)
Right	0.96 (0.76-0.99)	0.64 (0.37-0.95)	1.00 (0.98-1.00)	0.96 (0.88-0.99)	0.92 (0.46-0.99)	0.90 (0.37-0.98)	0.95 (0.89-0.97)	0.91 (0.72-0.97)
Female	0.91 (0.81-0.98)	0.49 (0.36-0.75)	0.99 (0.98-1.00)	0.93 (0.85-0.99)	0.87 (0.72-0.97)	0.80 (0.56-0.94)	0.99 (0.92-1.00)	0.98 (0.71-0.99)
Male	0.86 (0.78-0.93)	0.37 (0.26-0.61)	1.00 (0.98-1.00)	0.99 (0.97-1.00)	0.97 (0.62-0.99)	0.26 (0.03-0.99)	0.97 (0.78-0.99)	0.45 (0.13-0.91)

eTable 8. Pairwise intraclass correlation coefficient (ICC) and 95% confidence intervals, measuring the agreement of the volumes segmented, between graders (intergrader) and between the model and graders for each scan (model-grader).

Dice Similarity Coefficients

	Intraretinal fluid			Subretinal fluid			Subretinal hyperreflective material			Pigment epithelial detachment		
	n	Intergrader	Model-Grader	n	Intergrader	Model-Grader	n	Intergrader	Model-Grader	n	Intergrader	Model-Grader
All Scans	121	0.56 (0.40–0.69)	0.43 (0.29–0.66)	67	0.8 (0.66–0.87)	0.78 (0.57–0.85)	32	0.65 (0.37–0.78)	0.48 (0.11–0.65)	105	0.69 (0.56–0.78)	0.57 (0.34–0.70)
Set 1: Topcon-AMD	15	0.73 (0.59–0.77)	0.7 (0.39–0.76)	14	0.86 (0.80–0.89)	0.84 (0.72–0.86)	N/A	N/A	N/A	13	0.74 (0.50–0.79)	0.62 (0.41–0.85)
Set 2: Topcon-AMD	17	0.44 (0.32–0.53)	0.37 (0.14–0.57)	25	0.81 (0.73–0.87)	0.8 (0.74–0.87)	17	0.64 (0.27–0.78)	0.44 (0.07–0.67)	46	0.66 (0.54–0.77)	0.62 (0.46–0.76)
Set 2: Topcon-DME	41	0.59 (0.48–0.69)	0.51 (0.30–0.67)	3	0.78 (0.75–0.85)	0.45 (0.11–0.78)	N/A	N/A	N/A	N/A	N/A	N/A
Set 2: Heidelberg-AMD	25	0.36 (0.20–0.56)	0.36 (0.24–0.55)	20	0.68 (0.53–0.79)	0.63 (0.35–0.70)	15	0.65 (0.48–0.77)	0.49 (0.14–0.62)	46	0.69 (0.60–0.79)	0.42 (0.25–0.60)
Set 2: Heidelberg-DME	23	0.62 (0.52–0.73)	0.39 (0.31–0.61)	5	0.82 (0.76–0.90)	0.81 (0.68–0.83)	N/A	N/A	N/A	N/A	N/A	N/A

Table 9. Median and interquartile range of Dice Similarity Coefficients for scans where the feature was considered present across all 3 segmentations.

	Intraretinal fluid			Subretinal fluid			Subretinal hyperreflective material			Pigment epithelial detachment		
	n	Intergrader	Model-Grader	n	Intergrader	Model-Grader	n	Intergrader	Model-Grader	n	Intergrader	Model-Grader
Topcon	73	0.57 (0.44–0.71)	0.52 (0.30–0.69)	42	0.82 (0.77–0.89)	0.82 (0.68–0.86)	17	0.64 (0.27–0.78)	0.44 (0.07–0.67)	59	0.68 (0.52–0.78)	0.62 (0.45–0.77)
Heidelberg	48	0.52 (0.33–0.67)	0.38 (0.28–0.59)	25	0.7 (0.55–0.82)	0.64 (0.54–0.81)	15	0.65 (0.48–0.77)	0.49 (0.14–0.62)	46	0.69 (0.60–0.79)	0.42 (0.25–0.60)
Heidelberg 25 B-scans	33	0.54 (0.33–0.66)	0.35 (0.25–0.49)	14	0.68 (0.55–0.81)	0.63 (0.52–0.72)	9	0.64 (0.34–0.77)	0.43 (0.22–0.58)	27	0.73 (0.64–0.80)	0.46 (0.26–0.63)
Heidelberg 49 B-scans	15	0.49 (0.36–0.65)	0.51 (0.33–0.65)	11	0.74 (0.56–0.85)	0.66 (0.56–0.84)	6	0.69 (0.58–0.78)	0.55 (0.02–0.70)	19	0.62 (0.57–0.75)	0.38 (0.24–0.56)
AMD (all)	57	0.46 (0.31–0.67)	0.41 (0.27–0.66)	59	0.8 (0.66–0.87)	0.78 (0.57–0.85)	32	0.65 (0.37–0.78)	0.48 (0.11–0.65)	10	0.69 (0.56–0.78)	0.57 (0.34–0.70)
AMD (mild)	11	0.33 (0.10–0.57)	0.3 (0.12–0.45)	4	0.43 (0.18–0.66)	0.39 (0.03–0.69)	3	0.34 (0.23–0.45)	0.15 (0.02–0.33)	38	0.67 (0.55–0.74)	0.45 (0.28–0.59)
AMD (moderate)	16	0.43 (0.31–0.54)	0.38 (0.27–0.54)	17	0.73 (0.64–0.82)	0.71 (0.54–0.84)	13	0.64 (0.38–0.75)	0.41 (0.05–0.61)	28	0.7 (0.60–0.77)	0.6 (0.37–0.69)
AMD (severe)	30	0.53 (0.35–0.73)	0.53 (0.30–0.73)	38	0.82 (0.75–0.89)	0.8 (0.63–0.86)	16	0.73 (0.51–0.80)	0.59 (0.35–0.68)	39	0.72 (0.53–0.85)	0.63 (0.39–0.80)
AMD (baseline)	35	0.53 (0.36–0.72)	0.52 (0.28–0.72)	36	0.81 (0.66–0.89)	0.78 (0.59–0.85)	18	0.6 (0.36–0.78)	0.49 (0.25–0.68)	49	0.64 (0.55–0.76)	0.56 (0.37–0.66)
AMD (3m)	11	0.36 (0.27–0.58)	0.41 (0.28–0.65)	11	0.78 (0.51–0.84)	0.8 (0.27–0.86)	9	0.64 (0.16–0.77)	0.39 (0.02–0.58)	29	0.7 (0.49–0.83)	0.54 (0.35–0.73)
AMD (12m)	11	0.32 (0.30–0.47)	0.33 (0.21–0.41)	12	0.8 (0.72–0.83)	0.75 (0.61–0.85)	5	0.72 (0.64–0.75)	0.47 (0.17–0.68)	27	0.7 (0.57–0.78)	0.58 (0.28–0.76)
DME (all)	64	0.6 (0.49–0.70)	0.47 (0.31–0.66)	8	0.8 (0.75–0.90)	0.76 (0.57–0.84)	N/A			N/A		
DME (mild)	26	0.61 (0.45–0.68)	0.48 (0.32–0.61)	2	0.6 (0.54–0.66)	0.3						
DME (moderate)	21	0.62 (0.54–0.77)	0.54 (0.34–0.71)	1	0.76 (0.76–0.76)	0.83 (0.82–0.83)						
DME (severe)	17	0.57 (0.54–0.70)	0.34 (0.27–0.53)	5	0.9 (0.82–0.92)	0.81 (0.68–0.88)						
DME (baseline)	24	0.57 (0.48–0.69)	0.38 (0.30–0.67)	3	0.78 (0.77–0.84)	0.69 (0.51–0.80)						
DME (3m)	23	0.61 (0.56–0.70)	0.45 (0.31–0.61)	4	0.87 (0.79–0.92)	0.84 (0.60–0.88)						
DME (12m)	17	0.64 (0.44–0.71)	0.56 (0.32–0.68)	1	0.48 (0.48–0.48)	0.62 (0.61–0.63)						
Left	42	0.61 (0.49–0.70)	0.55 (0.30–0.70)	19	0.83 (0.73–0.88)	0.82 (0.43–0.86)						
Right	31	0.5 (0.35–0.69)	0.36 (0.30–0.68)	23	0.82 (0.79–0.89)	0.82 (0.73–0.87)	12	0.67 (0.36–0.79)	0.45 (0.11–0.72)	33	0.7 (0.54–0.77)	0.61 (0.50–0.77)
Female	60	0.57 (0.43–0.68)	0.45 (0.31–0.61)	35	0.82 (0.72–0.89)	0.79 (0.63–0.85)	17	0.64 (0.40–0.78)	0.44 (0.05–0.64)	57	0.7 (0.57–0.79)	0.59 (0.36–0.72)
Male	61	0.56 (0.35–0.69)	0.39 (0.25–0.68)	32	0.78 (0.63–0.85)	0.78 (0.49–0.85)	15	0.65 (0.33–0.77)	0.54 (0.22–0.68)	48	0.65 (0.54–0.76)	0.5 (0.32–0.66)

eTable 10. Median and interquartile range of DSC for scans where the feature was considered present across all 3 segmentations.