tvst

Article

Automated Quantification of Pathological Fluids in Neovascular Age-Related Macular Degeneration, and Its Repeatability Using Deep Learning

Irmela Mantel[1], Agata Mosinska[2], Ciara Bergin[1], Maria Sole Polito[1], Jacopo Guidotti[1], Stefanos Apostolopoulos[2], Carlos Ciller[2], and Sandro De Zanet[2]

[1] Department of Ophthalmology, University of Lausanne, Jules–Gonin Eye Hospital, Fondation Asile des Aveugles, Lausanne, Switzerland
[2] RetinAI Medical AG, Bern Switzerland

Correspondence: Irmela Mantel, Jules Gonin Eye Hospital, 15 Avenue de France, CP 5143, CH-1000 Lausanne 2, Switzerland. e-mail: irmela.mantel@fa2.ch

Received: May 26, 2020
Accepted: January 4, 2021
Published: April 16, 2021

Keywords: age-related macular degeneration; exudation; optical coherence tomography; fluid quantification; deep learning algorithm

Citation: Mantel I, Mosinska A, Bergin C, Polito MS, Guidotti J, Apostolopoulos S, Ciller C, De Zanet S. Automated quantification of pathological fluids in neovascular age-related macular degeneration, and its repeatability using deep learning. Trans Vis Sci Tech. 2021;10(4):17, https://doi.org/10.1167/tvst.10.4.17

Purpose: To develop a reliable algorithm for the automated identification, localization, and volume measurement of exudative manifestations in neovascular age-related macular degeneration (nAMD), including intraretinal (IRF), subretinal fluid (SRF), and pigment epithelium detachment (PED), using a deep-learning approach.

Methods: One hundred seven spectral domain optical coherence tomography (OCT) cube volumes were extracted from nAMD eyes. Manual annotation of IRF, SRF, and PED was performed. Ninety-two OCT volumes served as training and validation set, and 15 OCT volumes from different patients as test set. The performance of our fluid segmentation method was quantified by means of pixel-wise metrics and volume correlations and compared to other methods. Repeatability was tested on 42 other eyes with five OCT volume scans acquired on the same day.

Results: The fully automated algorithm achieved good performance for the detection of IRF, SRF, and PED. The area under the curve for detection, sensitivity, and specificity was 0.97, 0.95, and 0.99, respectively. The correlation coefficients for the fluid volumes were 0.99, 0.99, and 0.91, respectively. The Dice score was 0.73, 0.67, and 0.82, respectively. For the largest volume quartiles the Dice scores were >0.90. Including retinal layer segmentation contributed positively to the performance. The repeatability of volume prediction showed a standard deviations of 4.0 nL, 3.5 nL, and 20.0 nL for IRF, SRF, and PED, respectively.

Conclusions: The deep-learning algorithm can simultaneously acquire a high level of performance for the identification and volume measurements of IRF, SRF, and PED in nAMD, providing accurate and repeatable predictions. Including layer segmentation during training and squeeze-excite block in the network architecture were shown to boost the performance.

Translational Relevance: Potential applications include measurements of specific fluid compartments with high reproducibility, assistance in treatment decisions, and the diagnostic or scientific evaluation of relevant subgroups.

Introduction

Age-related macular degeneration (AMD) is a disorder with a high prevalence of around 170 million people affected globally.[1] Its neovascular form (nAMD) is characterized by pathological fluid exudation from neovascularization. Current treatment relies on repetitive intravitreal injections of antivascular endothelial growth factor (anti-VEGF). However, the key to achieving the best possible visual outcome is the early discovery of nAMD, prompt treatment initiation, and adequate retreatment strategy. A pathological fluid discovery on spectral domain optical coherence tomography (SD-OCT) is currently considered the most sensitive noninvasive imaging technique for both

screening and as a retreatment criterion.[2] However, because of a high number of patients at risk and those requiring long-term repetitive retreatment, health care systems are overwhelmed by the medical need. In clinical practice, the evaluation of the presence or absence of pathological fluid biomarkers in OCT volumes is a crucial requirement. Volume measurements of the fluid would be just as interesting and helpful; however, these measures are not available in the OCT device software, and manual quantification is too time consuming. Thus objective quantification relies on central retinal thickness measures, both in clinical patient care and in clinical research.

Computer-based automated image interpretation is a valuable clinical and research tool, and it has the potential to simplify patient care in terms of screening and follow-up. Several studies have shown the feasibility of deep-learning methods in ophthalmology[3–7] and in particular, for fluid detection.[4,8–11] Deep learning represents a data-driven approach, where features useful for describing the problem at hand are automatically discovered by processing a large amount of annotated data. This contrasts with previous heuristic-based approaches, which relied on subjective hand-designed features. Fully convolutional neural network (FCNN)—a type of deep-learning approach—is particularly suited for the segmentation task as it can classify every image pixel (for example the fluid class) considering the surrounding visual context.

The challenges for automated image evaluation are not only the determination of the presence or absence of pathological lesions and the precision, sensitivity, and specificity of the segmentation, but also the classification of fluid and the localization in the three dimensions with respect to the retinal layers and central foveal area. Thus far, important advances have been made without integrating all of these challenges.[4,8–10,12] In addition, one of the known issues of deep learning methods is its instability and dependence on the input distribution.[13,14] However, to the best of our knowledge, the reproducibility of volumetric algorithm results has not been tested so far.

The goal of this study was to develop a fully automated segmentation algorithm based on supervised deep-learning to detect, quantify, and localize the amount of pathological fluid in the SD-OCT in nAMD. In addition, the reproducibility of the volumetric fluid measurements was to be tested. Our approach integrated all three typical fluid compartments simultaneously, including intraretinal, subretinal, and subretinal pigment epithelium spaces, aiming not only to detect these fluid types, but also to predict its localization and quantity. Our additional aim was to compare our segmentation approach to other state-of-the-art techniques and to evaluate the influence of retinal layer segmentation, as well as parts of neural network architecture on the overall performance. Thus this study contributes to the development of the high-performing algorithm for volumetric measurements of pathological retinal fluid, including for the first time repeatability measures.

Methods

Patient Selection

For algorithm training and testing, the SD-OCT data set 1 was extracted from a consecutive series of two prospective study data. The studies were designed to test the safety and efficacy of an observe-and-plan regimen using aflibercept and ranibizumab respectively for the treatment of naïve nAMD.[1] At baseline, a full ophthalmic assessment with best corrected visual acuity, slit lamp examination, intraocular pressure, dilated fundus examination, SD-OCT examination, fluorescein angiography, and indocyanine green angiography was performed. An experienced retinal specialist (I.M.) confirmed the diagnosis of nAMD and its exudative activity before inclusion into the study. Patients were allowed into the original study with image quality sufficient for determining the presence of pathological fluid. However, no high-quality imaging was required.

The OCT data set 1 consisted of 107 SD-OCT volumes (49 b-scans, 6 × 6 mm cube examination), extracted from the Heidelberg Spectralis device (Heidelberg Engineering, Heidelberg, Germany). Forty-nine SD-OCT volumes (46%) corresponded to a treatment naïve situation from baseline of the observe-and-plan study (aflibercept study). An additional 58 SD-OCT volumes (54%) were selected from a follow-up visit.

The observe-and-plan study and the present post-hoc image evaluation adhered to the tenets of the Declaration of Helsinki, and the protocols were approved by the regional ethics committee (Swissethics Vaud 22/13, and 2017/00493). Patients had given written informed consent.

A second data set (data set 2) was acquired from a prospective study (protocol number CER-VD 2017-02175), aiming to document the reproducibility of the algorithm volumetric measures on repeated OCT acquisition from the same eye on the same day. Data were acquired from 42 eyes of 40 patients (none of whom was treatment-naïve), recruited from routine

AMD clinic if intraretinal or subretinal fluid was present and who gave informed consent for five successive OCT volume acquisitions (49 lines, 6 × 6 mm cube) from the same eye, centered on the fovea, on the Heidelberg Spectralis machine.

Image Export and Preprocessing

Complete SD-OCT volumes were extracted from the Heidelberg Spectralis device. Images were encoded, and all patient identifiers were removed from data packages. Each OCT volume captured the area of 6 × 6 × 2 mm^3 centered on the fovea, including at least 49 B-scans (standard B-scan density in the observe-and-plan study). The complete set of B-scans was exported as an E2E file. The E2E files were exported into OmniViewer (http://omniviewer.io)—a dedicated software application enabling the annotation of structures in three-dimensional (3D) medical images.

Manual Annotation

Of the 107 OCT volumes of data set 1, 92 (49 patients, 49 eyes) were selected for algorithm training and validation, and 15 OCT volumes (15 patients, 15 eyes) were kept aside for testing. These two subsets had no overlap because they included different patients.

To ensure a large variability in the training data and simultaneously reduce the annotation effort, some of the selected OCT volumes were subsampled, and only a subset of 10 B-scans per volume (every fifth B-scan) was annotated. The manual annotation was performed by three experienced clinicians (I.M., J.G., M.S.P.), delineating the area of IRF, SRF, and PED in each selected B-scan. The following definitions were applied for manual segmentation:

Intraretinal fluid was defined as cystoid spaces within the neuroretina, hyporeflective in comparison with the surrounding retinal tissue. A minimum diameter of 25 µm was required. In case these cystic spaces contained some reflective material, they were segmented if a clear limit to the normal retinal tissue could be identified. Excluded were optically empty triangular spaces adjacent to the internal limiting membrane and outer retinal tubulations identified by their hyperreflective border around the cystoid formation located in the outer nuclear layer.

Subretinal fluid was defined as hyporeflective separation of the photoreceptor layer from the retinal pigment epithelium. Small reflective dots within the SRF could be included (<25 µm); however, larger areas of reflective material were not included. This aimed to avoid inclusion of neovascular tissue, hemorrhage, fibrosis, or pseudovitelliform material, or undefined deposits.

Pigment epithelium detachment was identified as dehiscence of the pigment epithelium from Bruch membrane. The entire area was segmented independently of its internal reflectivity. However, small elevations of less than 25 µm were not segmented (small and middle-size drusen).

Examples of fluid segmentation ground-truth are shown in Figure 1. In cases of uncertainty, an experienced retinal specialist (I.M.) decided on the segmentation with the best clinical relevance. This decision was based on the complete set of acquired OCT B-scans, using the continuity of the subsequent B-scans for layer definitions.

After the first 20 manually segmented OCT, the algorithm was pretrained with the B-scan and segmentation information. The next series first underwent an automated segmentation analysis, corrected by the human reader, and then it was fed back into the algorithm. Several loops of refined algorithm learning were performed.

Algorithm Development

Training Dataset

Our training dataset consisted of 2680 manually segmented B-scans (originating from 92 OCT volumes) of size 496 × 512 pixels each. Of these, 90% were used for training and 10% for validation (monitoring the learning process and selecting the best model). To add more variation to our data, we applied a standard set of image augmentations to the training procedure. Those included horizontal and vertical shifts, left-right flipping, rotations up to ±20°, scaling, Gaussian blur, contrast changes, and additive noise. The B-scans were resized to size 256 × 256 pixels, and the image intensity was normalized to the range 0.0 to 1.0.

Network

To obtain the segmentations, we trained an encoder/decoder style FCNN.[15] The input to the network was a B-scan, and the output was a corresponding probability map with C channels where C is the number of classes. Our network consists of a downsampling path that extracts image features at different levels of abstraction and an upsampling branch that synthesizes information from the feature maps to compute the final predictions. The network depth was set to 5, corresponding to four downsampling steps. The residual block used in our network comprised two convolutional layers with ReLu activations in between, with each layer having 22, 33, 44, 88, and 176 filters respective to increasing depth. On the

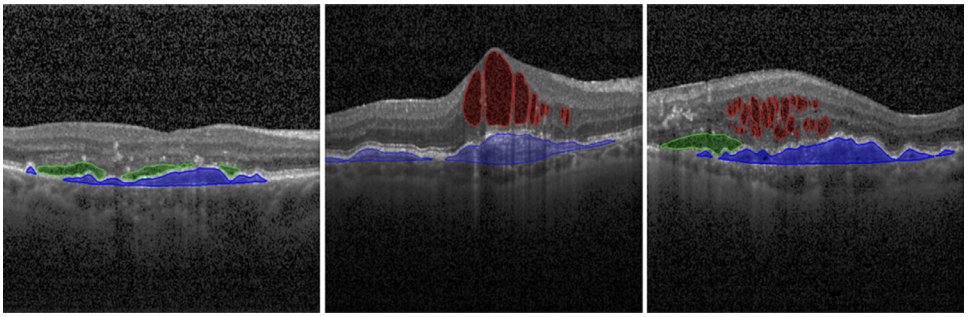

Figure 1. Examples of intraretinal exudation (*red*), subretinal fluid (*green*), and pigment epithelium detachment (*blue*) segmented manually by clinicians. Fluids present a large variability of appearances and scale, even within the same class, which makes automatic segmentation a particularly challenging task.

first level the convolutional kernel size was set to 7 × 7 to increase the context and on the subsequent levels it had size 3 × 3. Additionally, the two convolutional layers in the base block used atrous convolutions with a dilation factor of 1 in the first layer (standard convolutions) and 4 in the second layer to increase the receptive field. Finally, we combined the base block with a squeeze-excite block[16] at its end, which learns the importance of each feature channel. The schematic illustration of the network building block is shown in Figure 2.

All three fluids were predicted together, and we additionally used the presegmented retinal layer information to facilitate the learning process—we predicted six retinal layers jointly with the fluids and computed the loss on all nine classes; it was shown that adding this form of prior information helps with the segmentation,[8,11,17,18] and we hypothesize that it helps the method to keep the context. Layer presegmentation was obtained using the pretrained model of Apostolopoulos et al.[15] It was then combined with fluid manual segmentation, with fluid taking precedence. The final prediction was obtained by taking the softmax activation of the network output and assigning the class with highest probability to each pixel. We trained the network using our custom pipeline based on Pytorch, by optimizing the categorical cross-entropy for 120 epochs using the Adam optimizer[19] with an initial learning rate of 10^{-3}. We monitored the loss on the validation set and decreased the learning rate by 50% every time there was no improvement for seven epochs. The model with the lowest validation loss was saved for the final testing.

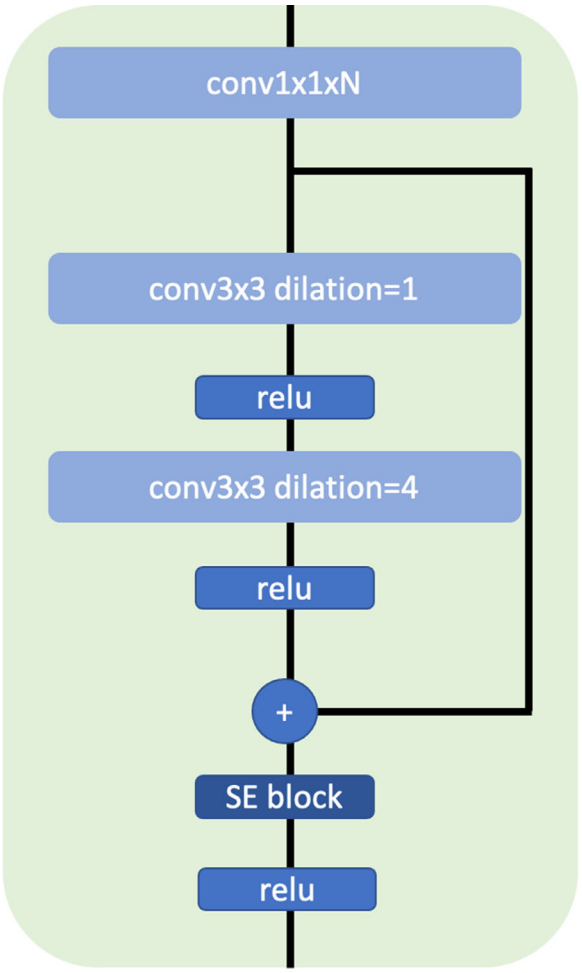

Figure 2. A schematic illustration of the building block of our network. Conv1×1 and conv3×3 refer to N-out channels convolutions with filter size 1×1 and 3×3, respectively, and SE block to squeeze-excite block.

Testing Datasets

(1) Fluid Detection and Quantification. To assess the viability of our algorithm, we tested it on the held-out

dataset. It consisted of 831 B-scans from 15 patients (15 OCT volumes), whose data was not used for training or validation. This setup allowed us to test the generalizability properties of our model. The algorithm performance was confronted with the manual segmentation of the test set. To evaluate the false-positive results in the detection task, we further included five scans of healthy patients.

(2) Reproducibility of Measurements. Dataset 2 (42 eyes with 5 OCT volumes each, same day, same OCT machine Spectralis) was processed with the fully developed algorithm, calculating the fluid volume for each of the three compartments. This was to ensure that the amount of fluid was the same, but each acquisition varied in terms of image quality, overall brightness, contrast, and alignment.

Evaluation Metrics

We evaluated our method in terms of fluid quantification and detection performance. To assess the former, we predicted fluid classes for each pixel in the test OCT volumes. We then computed classical segmentation metrics such as sensitivity, precision, and Dice score for scans that contained fluid. However, those metrics are usually biased by volumes that contain very little fluid; even small absolute errors have a considerable effect on the score. This makes the pixel-wise metrics difficult to compare between different methods and testing datasets. Therefore we estimated the fluid volumes for each testing case and compared them to the precise manually annotated volumes using two measures: Pearson's correlation coefficient and determination coefficient (R^2). The former measures the linear correlation between two variables (in our case, the true and predicted fluid volumes) and ranges between −1 to 1. The latter indicates the proportion of predicted fluid variation that can be explained by the annotated fluid volume and ranges between 0 to 1. The coefficients were computed in linear scale per whole volume and at different regions of the Early Treatment Diabetic Retinopathy Study (ETDRS) grid shown in Figure 1 of the supplementary material. The rationale behind it is that the fluids in the subfoveal region are particularly detrimental for visual acuity,[20] and thus mistakes in this zone should have higher penalties. We additionally used Bland-Altman plots to study bias in fluid volume prediction.

To compare our approach to other state-of-the-art segmentation methods (Unet, ReLayNet,[18] and RefNet[21]), we reimplemented them and trained and tested on our datasets. The utility of the proposed architecture improvements (dilated convolutions and squeeze-excite block), as well as the effect of layer information, was evaluated in an ablation study.

In addition, the fluid segmentation predictions and estimated volumes were used to classify the presence or absence of fluid in test OCT volumes and test B-scans. This served to investigate the performance of our model for pathology detection. The predicted segmentations were used to estimate the amount of fluid in every B-scan/OCT volume. This quantity was then thresholded to the presence or absence of each fluid type. The receiver operating characteristic curve, which depicts sensitivity plotted versus specificity for different thresholds of fluid volumes, was then computed along with area under the curve (AUC).

Finally, we performed a reproducibility study with data set 2 by predicting IRF, SRF, and PED and calculating total volumes for each fluid type. We then computed mean predicted fluid volume and standard deviation across 5 acquisitions, individually for each patient.

Results

Our model training took two days on a single GeForce GTX 1080 GPU (Nvidia, Santa Clara, CA, USA).

Evaluation of Fluid Quantification

The distribution of fluids in dataset 1 used for training the algorithm is presented in Table 1 and for evaluation of fluid quantification and detection in Table 2. In Table 4, we present pixel-wise segmentation metrics: sensitivity, precision, and Dice score. The latter represents a combination of sensitivity and precision, and thus it cumulates the impact of any error, whether false-positive or false-negative. The results showed the highest performance for PED (Dice score 0.819), middle for IRF (Dice score 0.728), and the lowest for SRF (Dice score 0.674). All of these outcome measures are in relation to the total lesion size, because the error is given as a percentage of the total. Furthermore, we evaluated the Dice scores for each quartile of lesion sizes, for IRF, SRF, and PED correspondingly; the results are shown in Figure 3. The highest quartile achieved a median Dice score per B-scan of 0.93, 0.93, and 0.90 for IRF, SRF, and PED, respectively. However, in the smallest lesions, the Dice scores were low.

The comparison with previously reported artificial intelligence methods—Unet, ReLayNet[18] and RefNet[21] is summarized in Table 4. On our dataset,

Table 1. Training Dataset 1 Statistics According to Expert Segmentation

	IRF	SRF	PED
Number of OCT volumes containing fluid	40 (43%)	40 (43%)	86 (93%)
Number of B-scans containing the fluid	472	407	1369
Median fluid volume [nL] per OCT volume (Standard deviation/range)	0.0 (259.8/0.0–1795.2)	0.0 (263.6/0.0–1903.1)	71.0 (414.3/0.0–2151.5)

IRF, intraretinal fluid; SRF, subretinal fluid; PED, pigment epithelium detachment; OCT, optical coherence tomography.

Table 2. Test Dataset 1 Statistics According to Expert Segmentation

	IRF	SRF	PED
Number of OCT volumes containing fluid	10 (67%)	11 (73%)	15 (100%)
Number of B-scans containing fluid	209	210	649
Median fluid volume [nL] per OCT volume (standard deviation/range)	20.4 (394.7/0.0–1546.9)	4.71 (154.3/0.0–538.3)	537.4 (272.3/64.9–1192.3)

IRF, intraretinal fluid; SRF, subretinal fluid; PED, pigment epithelium detachment; OCT, optical coherence tomography.

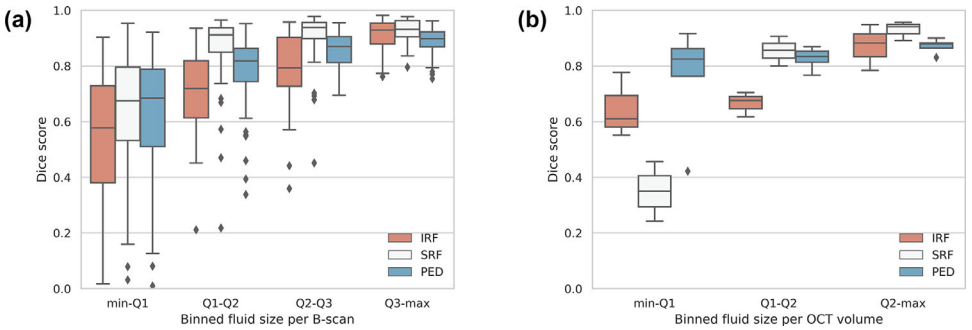

Figure 3. Dice score distribution over the total amount of fluids per B-scan (**A**)/optical coherence tomography (OCT) volume (**B**). Q1, Q2, and Q3 are quartiles of the data. Dice score measures the ratio of the overlap between the automatic segmentation and the manual annotation. B-scans/OCT volumes with a small amount of fluid tend to have lower scores, because even small mistakes of a few pixels may have a big effect on the metric. Note that the bin ranges of pigment epithelium detachment (PED) are bigger than that for intraretinal fluid (IRF) and subretinal fluid (SRF), because the test cases contain significantly more PED (see Table 2).

our method outperformed these reported methods for all types of fluids in terms of Dice scores. ReLayNet scored higher for precision of IRF and SRF, but its sensitivity was significantly worse than ours. RefNet achieved the highest precision for IRF, whereas IRF sensitivity dropped. Altogether, RefNet achieved comparable SRF and PED Dice scores to our method and lower for IRF.

Furthermore, we performed an ablation study to investigate the influence of squeeze-excite blocks and atrous convolutions, including layer information. The quantitative results are shown in Table 4.

As can be seen, adding squeeze-excite blocks improves the Dice scores for all fluids. A similar observation was made for addition of layer segmentation. Adding additional atrous convolutions (with dilation

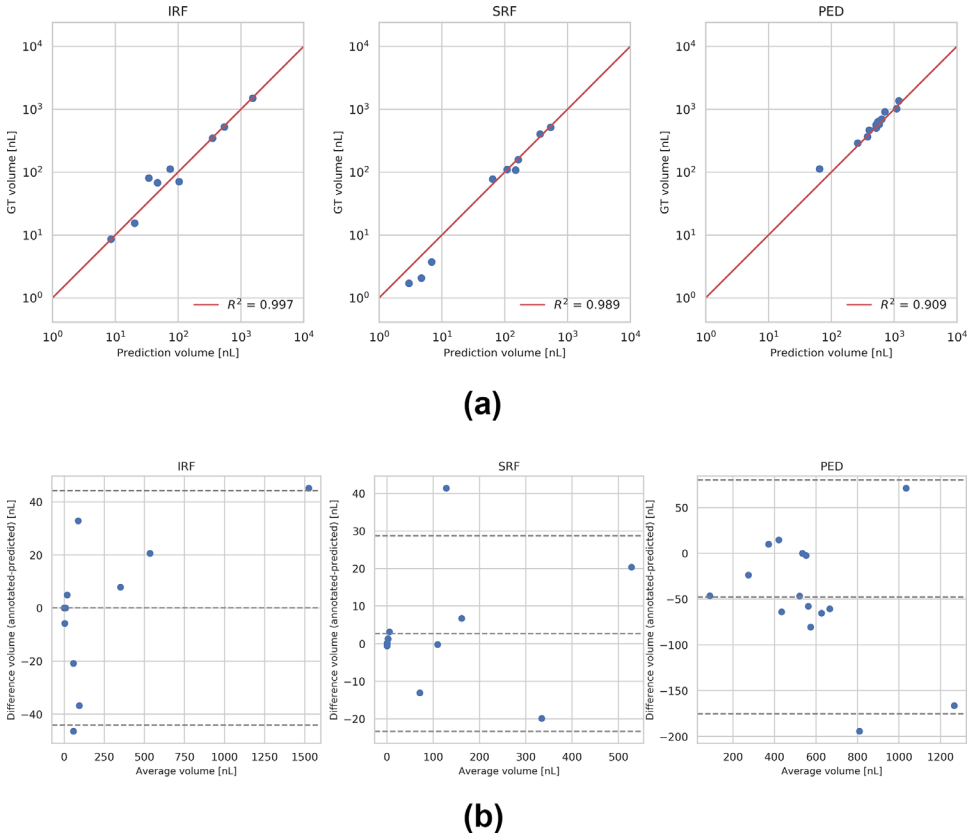

Figure 4. (**A**) Correlation of predicted and true fluid volumes in test optical coherence tomography volumes. The ideal correlation is shown by the *red line*. The predicted fluid volumes show very good correlation with volumes provided by the clinicians irrespective of the amount of fluid with Pearson's correlation coefficients of over 0.98 for every fluid type. (**B**) Bland-Altman plot shows agreement between predicted and annotated fluid volumes.

rate 4) improved the overall Dice scores for IRF and PED but decreased the performance for SRF.

Correlation analysis of fluid quantities detected by our model compared with that detected via expert annotation is shown in Figure 4A. The ideal correlation is represented by the red line and the measurements corresponding to each test OCT volume by blue dots. The Pearson correlation coefficient for IRF, SRF, and PED was 0.999, 0.996 and 0.976, respectively. Figure 4B presents Bland-Altman plot for full fluid volume prediction. We can deduce from it that the mean volume error for SRF and IRF is close to 0 and around 50nL for PED, which our method tends to oversegment.

The predicted fluid volume correlated well with the manually annotated quantity in all regions of ETDRS grid (Table 5). The distribution of ground-truth volumes matched the predicted distribution, as shown in Figure 5. Several examples of our predictions and comparison to the ground-truth are presented in Figure 6.

Evaluation of Fluid Detection

The prediction performance of fluid presence or absence is shown in Figure 7. The AUC for fluid detection receiver operating characteristic curve per B-scan (Fig. 7) was 0.97, 0.95, and 0.99 for IRF, SRF, and PED, respectively. The false-negative and false-positive results were mostly in the smallest volume quartile. The AUC for fluid detection per OCT volume was 1.0 for all three fluids, meaning that it was possible to find a fluid volume threshold in the prediction that can ideally separate healthy from nAMD volume scans.

Evaluation of Reproducibility

The distribution of fluids in the reproducibility dataset is presented in Table 3. The mean standard deviation for IRF, SRF, and PED was 4.0 nL, 3.5 nL, and 20.0 nL, which amounts to 6.9%, 3.6%, and 2.4% of the mean fluid volume, respectively. Figure 8

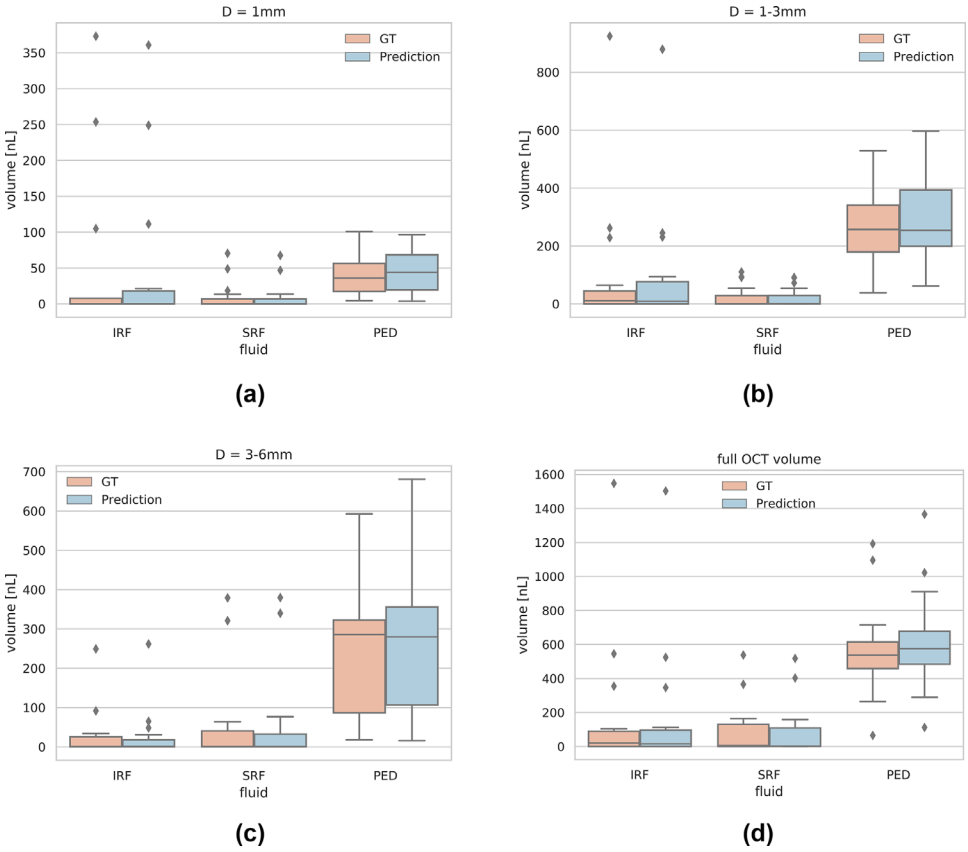

Figure 5. Distribution of ground-truth and predicted fluid volumes in the different regions of the Early Treatment Diabetic Retinopathy Study grid. Distributions obtained from the annotated and predicted volumes are very close, including outliers (see for example intraretinal fluid [IRF] and subretinal fluid [SRF] for D = 1 mm).

Table 3. Reproducibility Dataset 2 Statistics According To Automated Prediction

	IRF	SRF	PED
Number of OCT volumes containing fluid	182	179 (85%)	210 (100%)
Number of B-scans containing fluid	2781 (19%)	3943 (26%)	10182 (68%)
Median fluid volume [nL] per OCT volume (standard deviation/range)	5.6 (127.9/0.0-664.5)	16.6 (157.5/0.0-802.7)	478.0 (931.8/84.6-4424.4)

IRF, intraretinal fluid; SRF, subretinal fluid; PED, pigment epithelium detachment; OCT, optical coherence tomography.

presents a distribution of predicted fluid volumes for each eye in the reproducibility dataset. An example of a reproducibility measurement for one of the patients is presented in Supplementary Figure S2. The distribution of standard deviations of predicted volumes across 42 eyes is shown in Supplementary Figure S3.

Discussion

The presence of pathological fluid in nAMD is clinically directly relevant for the evaluation of its exudative activity and indirectly for its risk of further progression and visual loss. It is a treatment guiding biomarker,

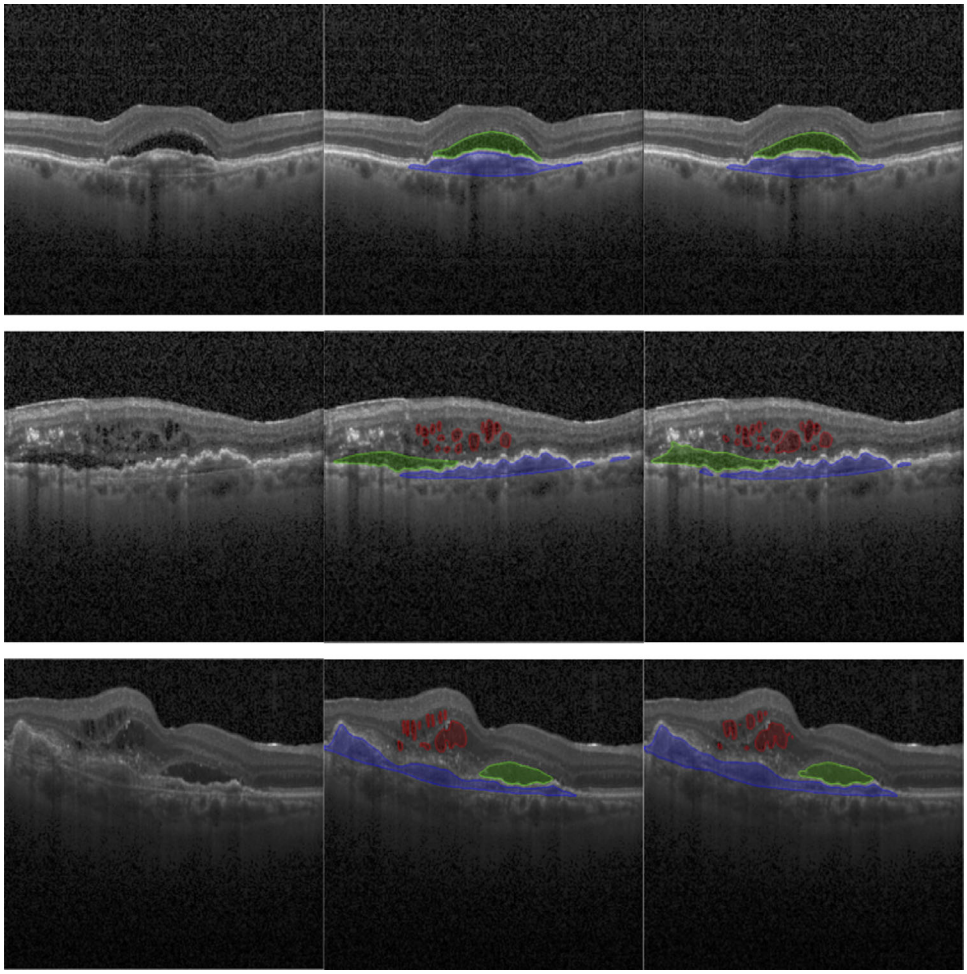

Figure 6. Examples of manual and automatic segmentations. From left to right: B-scan, ground-truth, and automatic prediction. Intraretinal fluid (*red*), subretinal fluid (*green*), and pigment epithelium detachment (*blue*).

and it is important for screening as well. Owing to the increasing ubiquity of OCT devices, more frequent and detailed examinations are now possible. However, despite major advances in image acquisition, image analysis requires manual interpretation. In a clinical setting, this is a bottleneck and requires major human resources. Therefore there is an interest in developing automated analysis methods that have a potential to significantly speed up the decision-making process. In addition, machine-integrated software does not allow for the quantification of pathological fluid. However, this would be of interest, particularly to quantify the exudation and measure the degree of treatment response where incomplete.

Along with other recent publications,[9–12] the present study confirmed the feasibility of the fully automated evaluation of pathological IRF, SRF, and PED in nAMD using a deep-learning algorithm. The developed algorithm showed a satisfying performance in terms of Dice score, volume correlation, and discovery rate, for all three fluid categories. In addition to the evaluation of the presence or absence of the three fluid categories, our study evaluated fluid volume, its localization and reproducibility. Fluid volumes have so far been reported for IRF and SRF only.[10] Our finding confirmed the excellent prediction of fluid volumes for all three categories; further, this was the case for all regions of the ETDRS grid. Additionally, fluid predictions were highly reproducible.

The current clinical practice for treatment guidance and screening relies on the simple discovery of the presence or absence of IRF and SRF. Therefore this parameter has vital importance for any automated OCT reading. The performance of our algorithms in terms of discovery rate was excellent (1.0 for all fluids per OCT volume). The performance per OCT volume, which is per eye and per time point, is the clinically relevant decision maker parameter. A reliable automated reading of the SD-OCT volumes could be of great significance for efficient image

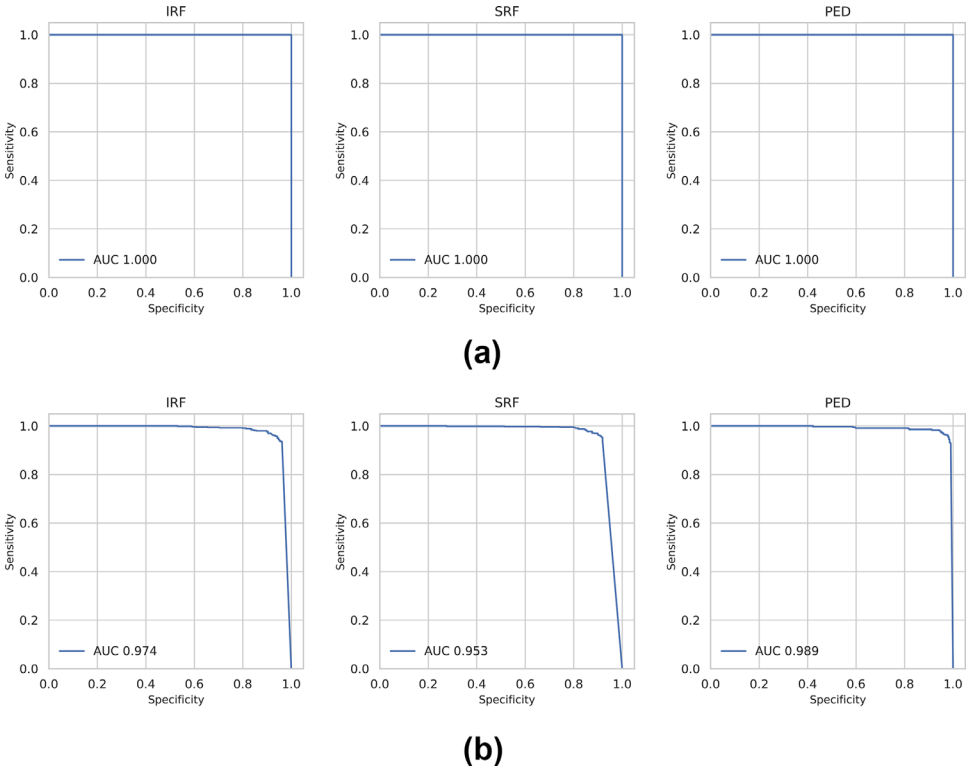

Figure 7. Fluid detection receiver operating characteristic in the optical coherence tomography (OCT) test volumes (**A**) and B-scans (**B**). It is obtained by classifying fluid presence/absence in every test OCT volume/B-scan by setting a threshold at different fluid volumes. Vertical axis presents sensitivity and horizontal 1-specificity.

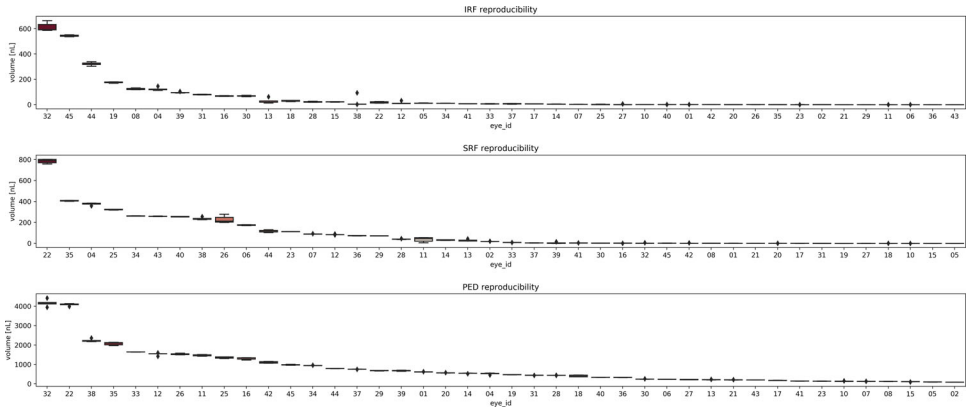

Figure 8. Boxplot results of reproducibility test for all eyes. Eyes are sorted according to fluid volume.

interpretation, allowing for the early discovery of nAMD, as well as the determination of the need for retreatment with anti-VEGF injections. The corresponding performance per B-scan (0.97, 0.95, and 0.99 for IRF, SRF, and PED) allows for appreciating the high discovery rate of fluid also per OCT slice. It has been shown that the presence of PED has a prognostic value, particularly in combination with IRF.[22] However, despite the high performance of the algorithm, important questions about liability remain if algorithms should indeed take over relevant decisions from clinicians. In this context, it might be interesting to consider that recent discussions suggest a lower importance for the presence or absence of SRF, at least in comparison to IRF and in low quantities.[20,23] Thus false-negative results in terms of subretinal fluid, more likely in small fluid quantities, may not be clinically relevant. In addition, it could be relevant to consider its location within or outside the central foveal area as it was done in the recent FLUID study.[23]

However, this is a more delicate approach, because the underlying neovascular disorder is a chronically progressive disease, and even extrafoveal activity may become visually threatening if progression is caused by insufficient treatment. Finally, the false-negative and -positive results in the discovery rate, leading to potential undertreatment or overtreatment if used for treatment decisions, are cases of small fluid volumes only. Thus deep-learning algorithms have become very close to being used for practical applications in clinics.

The volume metrics of fluid compartments are of increasing interest in clinical research about nAMD. Thus far, all large-scale studies used the readily available central retinal thickness and central retinal volume measures, which are inbuilt into the SD-OCT machines. However, these measures include normal tissue and pathological fluid, as well as any pathological mass such as the neovascular tissue, hemorrhage, or deposits. However, information about the different fluid compartments is lacking and requires time-consuming manual work. Measurements and changes in these compartments have several advantages: they represent only the exudative component and they are independent of the mass of retinal tissue or any pathological tissue. Such measurement would allow us to quantify the degree of response and recurrence. Therefore, the evaluation of the fluid compartments would be useful for clinical research, subgroup descriptions and potentially for patient care in the future. As mentioned above, it has recently been recognized that different fluid compartment may have a different relevance for treatment and prognosis: residual SRF may not have a negative influence on the visual acuity,[24,25] whereas IRF appears to be characterized by poorer visual outcomes yielding more intense treatment.[20,24] Moreover, PED tends to respond poorly to treatment, although an increase in volume predicts exudative activity into the retina.[26]

Our algorithm was able to provide highly precise volume metrics for each compartment of pathological fluid: correlation coefficients were above 0.90 for all three fluids. Such good correlation will allow for application in follow-up studies as well as to better quantify the effect of any intervention, according to the different fluid compartments. However, the described minor degree of measurement error will have to be integrated into the subsequent analysis. Dice score is a frequently used parameter to evaluate the performance of segmentation algorithms. It contains information about the overall similarity of the automated segmentation compared with the manual segmentation, confronting the true-positive area with both false-negative and false-positive areas. Our algorithm achieved comparable performance for all fluid classes—a mean volume Dice score of 0.728, 0.674, and 0.819, for IRF, SRF, and PED, respectively. Although the Dice score is the most commonly used measure to describe the performance of a segmentation algorithm, comparisons across different studies using different OCT material is difficult. In fact, the value of the Dice score depends on the size of the fluid: a few misclassified pixels in OCT volumes that contain little fluid tend to bias the Dice score toward lower values, even though the absolute error is very small. We computed the Dice scores for volume quartiles of each fluid. Not surprisingly, the larger fluid compartments (fourth volume quartile) performed much better, with the Dice score reaching 0.9 and higher for each type of fluid. Thus it is difficult to compare the performance of different algorithms using Dice scores, because the testing sets may contain very different quantities of fluid. We argue that using an additional, more clinically relevant metric—the correlation between the predicted and manually annotated fluid volumes—allows for better comparison between different methods.

To further examine the performance of our model, we evaluated the correlation metrics in different regions of the ETDRS grid, because it was shown that the subfoveal regions are particularly important for determining visual acuity.[20] Therefore we were most concerned with the mistakes in this area. Our approach achieved high values of both Pearson coefficient and R^2 in this region, as listed in Table 5.

The IRF showed the largest volume error relative to the total volume. This may be explained by the multiple small cystoid spaces with poorly defined borders and the B-scan images sometimes suggesting confluence. Another source of error is the instances of pockets created by the internal limiting membrane or epiretinal membranes, which are erroneously detected as IRF. This could be overcome by additional training including such scenarios. The PED volumes depend on the accurate detection of the Bruch membrane, which is sometimes challenging for the algorithm. More importantly, the RPE may sometimes be hidden in the neovascular mass, thus allowing mis-segmentation. For the SRF, the main challenge was found in the cases with reflective material, which is sometimes ill defined, particularly in treatment naïve cases. The examples of several failure cases are shown in Supplementary Figure S4. Bland-Altman plots (Fig. 4B) revealed that PED predictions showed the biggest bias leading to oversegmentation. According to our observations, this was mostly caused by PED predictions extending over Bruch's membrane in cases where it was not clearly visible or choroid presented high hypertransmission. In rare cases, parts of fibrosis and subretinal hyperreflective material were segmented as PED. Introducing

Table 4. Average Segmentation Metrics Per Test Volume in Dataset 1

	Sensitivity			Precision			Dice Score		
Method	IRF	SRF	PED	IRF	SRF	PED	IRF	SRF	PED
Unet	0.726	0.668	0.837	0.716	0.588	0.718	0.696	0.602	0.763
ReLayNet[18]	0.734	0.566	0.810	0.723	**0.829**	0.780	0.710	0.624	0.786
RefNet[21]	0.636	0.650	**0.875**	**0.796**	0.797	0.770	0.642	0.660	0.816
Ours without layer information	0.741	0.675	0.833	0.694	0.713	0.753	0.695	0.678	0.785
Ours without squeeze-excite	0.781	0.631	0.853	0.678	0.773	0.774	0.709	0.653	0.809
Ours without dilated convolutions	**0.838**	**0.704**	0.856	0.663	0.770	0.771	0.726	**0.720**	0.805
Ours (with layer information, squeeze-excite, and dilated convolutions)	0.778	0.647	0.860	0.721	0.732	**0.787**	**0.728**	0.675	**0.819**

IRF, intraretinal fluid; SRF, subretinal fluid; PED, pigment epithelium detachment.

Table 5. Correlation Coefficients of Predicted and Ground-Truth Test OCT Fluid Volumes in Different Regions of ETDRS Grid

Correlation Coefficient	ETDRS Grid Region	IRF	SRF	PED
R^2				
	1 mm	0.997	0.994	0.873
	1–3 mm	0.995	0.969	0.924
	3–6 mm	0.978	0.996	0.936
	Full OCT volume	0.997	0.989	0.909
Pearson's coefficient				
	1 mm	0.999	0.998	0.954
	1–3 mm	0.999	0.995	0.987
	3–6 mm	0.990	0.998	0.976
	Full OCT volume	0.999	0.996	0.976

ETDRS, Early Treatment Diabetic Retinopathy Study; IRF, intraretinal fluid; SRF, subretinal fluid; PED, pigment epithelium detachment; R^2, determination coefficient; OCT, optical coherence tomography.

additional classes that correspond to those pathologies could reduce this inconsistency and as a result also the bias.

Apart from computing the performance of our method, we investigated the effect of adding retinal layers information during training, as well as squeeze-excite block and dilated convolutions in the network architecture. Segmenting retinal layers along with fluids boosted the model performance. The additional classes corresponding to layers can be regarded as a loss regularizer and additionally help to infer fluid's anatomic location, which is especially important for distinguishing between fluid types. Squeeze-excite block further improved the performance.

However, atrous convolutions had a mixed effect on the segmentation. They improved the overall Dice scores for IRF and PED but decreased the performance for SRF. We hypothesize that dilated convolutions improve recall of larger objects and precision in noisy and ambiguous image regions at the expense of smaller objects recall, as they increase the effective receptive field. As shown in Table 4, the network trained with dilated convolutions reduced recall of IRF and SRF and increased recall of PED. In our test dataset both IRF and SRF generally manifested as smaller pockets than PED. At the same time dilated convolutions significantly increased the precision of IRF, which tends to be falsely detected in the noisy

image regions. As a result, the addition of dilated convolutions can be regarded as a trade-off between decreasing recall of small features and increasing recall of bigger features, as well as increasing precision in noisy image parts.

Our approach compared favorably to the baseline algorithms—Unet and ReLayNet. RefNet achieved the highest IRF precision among the compared methods and similar performance to our method for SRF and PED, but the Dice score of IRF was lower.

The reproducibility evaluation indicated that our algorithm produces stable predictions with respect to changes in the input images related to imaging variations that are often present in clinical practice. The mean standard deviation of IRF and SRF were within 4 nL. Higher mean standard deviation of PED (20.0 nL) can be explained by higher mean PED volume in the reproducibility set compared to SRF and IRF. To the best of our knowledge, this is the first reproducibility report of automated fluid volume measurement methods in the literature.

Our approach is based on processing two-dimensional (2D) B-scans. Although it is possible to extend it to process OCT volumes in 3D by replacing 2D convolutions by 3D version, we see two major benefits of the proposed 2D processing. First, our method is agnostic of the resolution between slices, which may differ significantly depending on the scan pattern. In some cases acquisitions also rely on a single B-scan or cross scans, in which case 3D processing would not be feasible. Second, given current limitations in training speed and performance, we selected a method that enabled capturing larger intersubject variability per training batch and allowed for processing larger B-scans. Finally, by constraining the model to a single B-scan, we also reduce the need for B-scans alignment before processing and achieve results that can be reliably compared in follow-up studies.

Deep-learning methods significantly outperform traditional machine learning methods based on handcrafted features; however, an increase in performance comes at a price. The success of data-driven approaches is highly dependent on the abundance of high-quality data, which is often a constraint. In addition, untreated nAMD shows a mixture of OCT signs, including not only pathological fluid but also hyperreflective materials in different layers, and severely disturbed retinal anatomy. Thus the detection and delineation of the pathological fluid may be particularly challenging. Moreover, accumulated fluid usually negatively influences the image quality, making this task even more challenging. This was a relevant challenge to our study, which included a high number of treatment naïve eyes (46%).

Previous studies focused on detecting one type of pathological fluid, mostly IRF causing macular edema.[9,11,12,17] Lee et al.[12] used a U-Net architecture to detect IRF with an average Dice score of 0.73. Contrary to our approach, the network operated on image patches instead of the entire B-scans, and the total prediction time was over 13 seconds per B-scan compared to 0.04 second in our case. Venhuizen et al.[11] showed that adding retinal segmentation prior helped with detecting intraretinal cystoid fluid, achieving an overall Dice score of 0.75. Roy et al.[18] and Asgari et al.[17] also showed that adding additional prior information in the form of retinal layers helped segmenting fluids in drusen and diabetic retinopathy. This corresponds well with our observation, and it is indeed not surprising because fluid definitions are closely related to the retinal layers. The detection of IRF and SRF in different retinal diseases was investigated by Schlegl et al.,[10] where the model obtained a precision of 0.78 and 0.81, and a sensitivity of 0.63 and 0.71, for IRF and SRF, respectively. Furthermore, they compared the distributions of predicted and annotated fluid volumes achieving an R^2 of 0.68 and 0.65 for IRF and SRF in Spectralis scans, and Pearson coefficients of 0.86 and 0.85, respectively. The multiclass fluid detector introduced by Lu et al.[8] and Lee et al.[9] are the closest to our method, because they distinguished simultaneously between IRF, SRF, and PED, obtaining Dice scores between 0.74 to 0.85 and 0.75 to 0.86, respectively. However, the method of Lu et al.[8] consisted of several refining steps, which increased the computational burden and required more parameter tuning. On the contrary, our method is trained in a single step, and it greatly reduces the processing time and complexity. In addition, these publications did not contain information about fluid volumes and their correlations.

Fluid segmentation is one of the well-studied problems in retinal image analysis. Despite this fact, several important points have not been considered so far (or discussed only partially) in the previous literature, which we discuss in this article. Those include the following:

- Integrating a reproducibility study, which enables assessment of reliability and stability of the automated prediction
- Joint prediction of all three types of fluids, while most of the published work tackles only one or two types of fluid
- Discussion of Dice score bias depending on the distribution of fluid volumes in the test set and adding an additional performance metric (correlation of the predicted and manually annotated fluid volumes) in different retinal regions, which

is more clinically relevant for fluid monitoring and patient progression tracking than standard computer vision metrics. To the best of our knowledge, fluid size bias has not been studied in depth in the literature for OCT fluid segmentation.

- Ablation study highlighting the importance and effect of squeeze-excite blocks and dilated convolutions

We believe that those contributions are important for assessing feasibility of automatic fluid segmentation in everyday clinical practice.

The limitations of the present study include the relatively limited number of included OCT volumes, and the restriction to our own center. In addition, our method was tested only on OCT scans acquired with the Heidelberg Spectralis machine. A potential follow-up study should include data from other devices to investigate the generalizability to other vendors. Images with very poor quality were excluded before this study. We did not perform any post hoc selection of images based on their quality; therefore we can assume that the included cases are relatively representative for clinical situations. A larger validation dataset originating from multiple centers would allow for a large-scale evaluation.

The potential of the automated detection of pathological fluid and its attribution to the different compartments of the retina is large. The detection of early disorder and individualized retreatment guidance are the most evident benefits. The precise localization of the fluid, as well as its configuration in three dimensions can further help differential diagnosis, evaluation of individual treatment response, and identification of associated factors. The quantification of the fluid in its compartment would allow for more-precise follow-up with patients and for more relevant measurement of change to the pathological compartment, which contrasts with the overall measure of central retinal thickness often used in clinical trials. In conclusion, the precision of deep-learning algorithms for the identification and segmentation of pathological IRF and SRF, as well as PED, has reached a very promising level of performance, allowing for applications in research and clinical activities in the near future.

Acknowledgments

Supported by a research grant from a fund dedicated for research in age-related macular degeneration.

Disclosure: **I. Mantel,** None; **A. Mosinska,** None; **C. Bergin,** None; **M. S. Polito,** None; **J. Guidotti,** None; **S. Apostolopoulos,** None; **C. Ciller,** None; **S. De Zanet,** None

References

1. Parvin P, Zola M, Dirani A, Ambresin A, Mantel I. Two-year outcome of an observe-and-plan regimen for neovascular age-related macular degeneration treated with Aflibercept. *Graefes Arch Clin Exp Ophthalmol*. 2017;255:2127–2134.
2. 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;31:26–44.
3. Abràmoff MD, Lou Y, Erginay A, et al. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Invest Ophthalmol Vis Sci*. 2016;57:5200–5206.
4. de Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*. 2018;24:1342–1350.
5. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316:2402–2410.
6. Lee CS, Baughman DM, Lee AY. Deep learning is effective for classifying normal versus age-related macular degeneration OCT images. *Ophthalmol Retina*. 2017;1:322–327.
7. Kurmann T, Yu S, Márquez-Neila P, et al. Expert-level automated biomarker identification in optical coherence tomography scans. *Sci Rep*. 2019;9:13605.
8. Lu D, Heisler M, Lee S, et al. Deep-learning based multiclass retinal fluid segmentation and detection in optical coherence tomography images using a fully convolutional neural network. *Med Image Anal*. 2019;54:100–110.
9. Lee H, Kang KE, Chung H, Kim HC. Automated segmentation of lesions including subretinal hyperreflective material in neovascular age-related macular degeneration. *Am J Ophthalmol*. 2018;191:64–75.
10. Schlegl T, Waldstein SM, Bogunovic H, et al. Fully automated detection and quantification of macular fluid in OCT using deep learning. *Ophthalmology*. 2018;125:549–558.

11. 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. *Biomedical Optics Express*. 2018;9:1545–1569.
12. Lee CS, van Ginneken B, Liefers B, et al. Deep-learning based, automated segmentation of macular edema in optical coherence tomography. *Biomedical Optics Express*. 2017;8:3440–3448.
13. Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572. 2014 Dec 20.
14. Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199. 2013 Dec 21.
15. Apostolopoulos S, de Zanet S, Ciller C, Wolf S, Sznitman R. Pathological OCT retinal layer segmentation using branch residual U-shape networks. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2017:294–301.
16. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2018;7132–7141.
17. Asgari R, Orlando JI, Waldstein S, et al. Multiclass segmentation as multitask learning for drusen segmentation in retinal optical coherence tomography. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2019:192–200.
18. Roy AG, Conjeti S, Karri SPK, et al. ReLayNet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks. *Biomed Opt Express*. 2017;8:3627–3642.
19. Kingma DP, Ba JL. Adam: A method for stochastic optimization. In: *International Conference on Learning Representations*. 2015:1–15.
20. Jaffe GJ, Martin DF, Toth CA, et al. Macular morphology and visual acuity in the comparison of age-related macular degeneration treatments trials. *Ophthalmology*. 2013;120:1860–1870.
21. Guo Y, Hormel TT, Xiong H, Wang J, Hwang TS, Jia Y. Automated segmentation of retinal fluid volumes from structural and angiographic optical coherence tomography using deep learning. *Transl Vis Sci Technol.* 2020;9(2):54.
22. Schmidt-Erfurth U, Waldstein SM, Deak GG, Kundi M, Simader C. Pigment epithelial detachment followed by retinal cystoid degeneration leads to vision loss in treatment of neovascular age-related macular degeneration. *Ophthalmology*. 2015;122:822–832.
23. Guymer RH, Markey CM, McAllister IL, et al. Tolerating Subretinal fluid in neovascular age-related macular degeneration treated with ranibizumab using a treat-and-extend regimen: FLUID Study 24-month results. *Ophthalmology*. 2019;126:723–734.
24. Jang L, Gianniou C, Ambresin A, Mantel I. Refractory subretinal fluid in patients with neovascular age-related macular degeneration treated with intravitreal ranibizumab: visual acuity outcome. *Graefes Arch Clin Exp*. 2015;253:1211–1216.
25. Wickremasinghe SS, Janakan V, Sandhu SS, Amirul-Islam FM, Abedi F, Guymer RH. Implication of recurrent or retained fluid on optical coherence tomography for visual acuity during active treatment of neovascular age-related macular degeneration with a treat-and-extend protocol. *Retina*. 2016;36:1331–1339.
26. Fragiotta S, Rossi T, Cutini A, Grenga PL, Vingolo EM. Predictive factors for development of neovascular age-related macular degeneration: a spectral-domain optical coherence tomography study. *Retina*. 2018;38:245–252.