Title

VALIDATION OF A DEEP LEARNING-BASED ALGORITHM FOR SEGMENTATION OF THE ELLIPSOID ZONE ON OPTICAL COHERENCE TOMOGRAPHY IMAGES OF AN USH2A-RELATED RETINAL DEGENERATION CLINICAL TRIAL

Journal

Retina, 42(7)

ISSN

0275-004X

Authors

Loo, Jessica
Jaffe, Glenn J
Duncan, Jacque L
et al.

Publication Date

2022-07-01

DOI

10.1097/iae.0000000000003448

Peer reviewed

HHS Public Access

Author manuscript

Retina. Author manuscript; available in PMC 2023 July 01.

Published in final edited form as:

Retina. 2022 July 01; 42(7): 1347–1355. doi:10.1097/IAE.0000000000003448.

Validation of a deep learning-based algorithm for segmentation of the ellipsoid zone on optical coherence tomography images of an *USH2A*-related retinal degeneration clinical trial

Jessica Loo, MEng[1],

Glenn J. Jaffe, MD[2],

Jacque L. Duncan, MD[3],

David G. Birch, PhD[4],

Sina Farsiu, PhD[1,2]

[1] Department of Biomedical Engineering, Duke University, Durham, NC, USA

[2] Department of Ophthalmology, Duke University Medical Center, Durham, NC, USA

[3] Department of Ophthalmology, University of California, San Francisco, San Francisco, CA, USA

[4] Retina Foundation of the Southwest, Dallas, TX, USA

Abstract

Purpose: To assess the generalizability of a deep learning-based algorithm to segment the ellipsoid zone (EZ).

Methods: The dataset consisted of 127 spectral-domain optical coherence tomography volumes from eyes of participants with *USH2A*-related retinal degeneration enrolled in the RUSH2A clinical trial (NCT03146078). The EZ was segmented manually by trained Readers and automatically by DOCTAD, a deep learning-based algorithm originally developed for macular telangiectasia type 2. Performance was evaluated using the Dice similarity coefficient (DSC) between the segmentations, and the absolute difference and Pearson's correlation of measurements of interest obtained from the segmentations.

Results: With DOCTAD, the average (mean ± SD, median) DSC was 0.79 ± 0.27, 0.90. The average absolute difference in total EZ area was 0.62 ± 1.41, 0.22 mm^2 with a correlation of 0.97. The average absolute difference in the maximum EZ length was 222 ± 288, 126 μm with a correlation of 0.97.

Correspondence: Sina Farsiu, PhD, 101 Science Drive, Campus Box 90281, Durham, NC 27705.

Financial disclosures: Jessica Loo (None), **Glenn J. Jaffe** (Consultant: Roche, EyePoint, Regeneron, Novartis, Adverum), **Jacque L. Duncan** (Consultant: ConeSight, DTx Pharma, Editas, Inc., Eyevensys, Gyroscope, Nacuity, ProQR Therapeutics, Inc, PYC Therapeutics, Inc., Relay, SparingVision, Spark Therapeutics, Inc., Vedere Bio. Funding: Acucela, Allergan/Abbvie, Biogen/Nightstarx, AGTC, ProQR Therapeutics, Stargazer, Foundation Fighting Blindness, Research to Prevent Blindness, National Eye Institute), **David G. Birch** (Consultant: AGTC, Nacuity, ProQR, Editas, Iveric, 4D Molecular Therapeutics, DTx Pharma. Funding: Biogen, AGTC, 4D Molecular Therapeutics, ProQR, Foundation Fighting Blindness, National Eye Institute EY09076), **Sina Farsiu** (None).

Conclusion: DOCTAD segmented EZ in *USH2A*-related retinal degeneration with good performance. The algorithm is potentially generalizable to other diseases and other biomarkers of interest as well, which is an important aspect of clinical applicability.

SUMMARY

We validate the clinical applicability and generalizability of DOCTAD, a deep learning-based algorithm originally developed for macular telangiectasia, to segment the ellipsoid zone on optical coherence tomography images of eyes with *USH2A*-related degeneration. The algorithm performed well on a diverse dataset from the large-scale, international, multi-center RUSH2A clinical trial (NCT03146078).

Keywords

Automatic; clinical applicability; deep learning; ellipsoid zone; generalizability; optical coherence tomography; segmentation; *USH2A*-related retinal degeneration

INTRODUCTION

Usher syndrome is an autosomal recessive genetic disease and the most common genetic cause of deaf-blindness.[1] Of the three forms of Usher syndrome, Usher syndrome type 2 (USH2) is the most common, accounting for over half of all Usher syndrome cases.[2] USH2 is characterized primarily by retinitis pigmentosa (RP) accompanied by mild to moderate hearing loss. USH2 is most commonly associated with variants in the *USH2A* gene, which is also the most common cause of non-syndromic autosomal recessive RP (i.e., without hearing loss).[3, 4]

USH2A-related RP is characterized by retinal photoreceptor ellipsoid zone (EZ) loss, representing macular photoreceptor degeneration.[5] Measurements of EZ loss, including length or area, are commonly used outcome measures in clinical studies of retinal diseases including macular telangiectasia type 2 (MacTel2).[6, 7] The change in total EZ area is one of the primary outcome measures in the ongoing, international, multi-center, longitudinal natural history study of participants with *USH2A*-related retinal degeneration (Rate of Progression of *USH2A*-related Retinal Degeneration (RUSH2A) Study, NCT03146078). Spectral domain optical coherence tomography (SD-OCT) provides non-invasive and high-resolution imaging of the retinal structures including the EZ, which can be measured quantitatively.

Automatic algorithms have been developed to analyze the EZ on SD-OCT images from a variety of retinal diseases.[8–20] While most of these algorithms were developed for a specific condition, generalizability is an important aspect of clinical applicability whereby the algorithm can be applied to other conditions. Trainable algorithms, such as modern deep learning-based algorithms, have especially strong potential to do so, given the appropriate training dataset. However, this is not necessarily always true. For example, we previously showed that a deep learning-based algorithm developed to segment exudative cystoid structures did not generalize well to segment degenerative cystoid structures in a different disease, thus warranting the development of a new algorithm.[21] Therefore, it is necessary

to explicitly validate algorithms for generalizability to other conditions. Besides broadening the scope of application of the algorithm, thereby increasing clinical applicability and potential for real-world clinical use, this generalizability also enables research efforts to be optimally focused on applications when existing algorithms do not suffice. Two notions of generalizability can be considered, which are (1) the generalizability of an existing methodology given the appropriate training dataset for a new application and (2) the generalizability of an existing model trained for a different application.

Another limitation to the clinical applicability of these automatic algorithms is that the performances are frequently reported on limited datasets with minimal heterogeneity. Therefore, while the reported performances of some of these techniques may be impressive, they do not reflect clinical applicability, and there is often a significant decline in performance when applied in real-world clinical settings. To minimize such discrepancies, we advocate for using complete clinical study datasets to ensure that the reported performance is as representative as possible of the performance in real-world clinical settings including diverse patient populations. Figure 1 shows some example images from the RUSH2A clinical study that exhibited a wide range of disease severity and macular findings. Such diversity increases segmentation complexity which may be difficult even for trained and experienced experts and would therefore provide a more robust evaluation of the algorithm's performance.

Deep OCT Atrophy Detection (DOCTAD) is a fully-automatic, deep learning-based, *en face* segmentation algorithm initially developed for and validated to segment EZ defects on SD-OCT images of eyes with MacTel2.[14, 22] One aspect of DOCTAD's clinical applicability has been previously validated, as the algorithm reproduced the statistically-significant expert-evaluated results in a phase 2 clinical trial for MacTel2.[22] In this article, we further assess DOCTAD's clinical applicability by validating its generalizability to segment the presence of EZ on SD-OCT images of *USH2A*-related RP from the RUSH2A clinical study. We focus on the generalizability of an existing methodology given the appropriate training dataset for a new application, as this approach would also be applicable for validating DOCTAD to segment biomarkers other than the EZ. However, since the EZ happens to be the biomarker of interest in both MacTel2 and *USH2A*-related RP, we also provide a brief analysis of the generalizability of an existing model trained for a different application.

METHODS

Dataset

The dataset consisted of OCT images of eyes of participants with *USH2A*-related RP enrolled in the RUSH2A clinical study (NCT03146078) at 16 clinical sites in North America, Europe, and the United Kingdom. The study was approved by the ethics board at each clinical site and adhered to the tenets of the Declaration of Helsinki. For each participant, the study eye was defined as the eye with better visual acuity at the baseline visit. All study eyes were imaged with Spectralis SD-OCT systems (Heidelberg Engineering, GmBH, Heidelberg, Germany) to obtain high resolution, macula-centered volume scans. All volumes were exported as raw binary files from Heidelberg Eye Explorer and converted to .tif images for image analysis.

Manual segmentation

Manual segmentation of the *en face* EZ was performed by a trained Reader using the Duke OCT Retinal Analysis Program (DOCTRAP V63.9, Duke University, Durham, North Carolina, USA) who labeled the absence or presence of EZ on each A-scan. Per the grading protocol, for each SD-OCT volume, the foveal B-scan on which the EZ was easiest to identify was segmented first, followed by the neighboring B-scans. If the absence or presence of EZ was unclear, the Reader used the assumption of EZ continuity from the fovea, to discourage the segmentation of small and discontinuous regions of EZ. A second, senior Reader reviewed the segmentations of the first Reader and made corrections when necessary. Figure 2 shows the manual segmentation process.

Automatic segmentation

Automatic segmentation of the *en face* EZ was performed by DOCTAD,[14, 22] a deep learning-based, *en face* segmentation algorithm. The convolutional neural network (CNN) architecture and training procedure were not modified in any way from the original publications. Briefly, the CNN was trained to classify clusters of A-scans with the absence or presence of EZ. The CNN architecture consisted of four blocks and three fully-connected layers followed by softmax activation in the final layer. Each block consisted of two convolutional layers, a batch normalization layer followed by rectified linear unit activation, and a max-pooling layer.

Six-fold cross-validation was used to train and test the automatic segmentation algorithm on all available data to avoid selection bias and to ensure independence between the training and testing sets. Participants were randomly divided into six groups of approximately equal size. Five groups were designated as the training set while the remaining group was designated as the testing set. The groups were then rotated such that each group was used once for testing. For validation, one group from the training set was designated as the hold-out validation set.

For training, clusters of A-scans (with dimensions 256 × 16 × 5) and their corresponding labels (0 for absence of EZ and 1 for presence of EZ in the central A-scan) were randomly extracted from the SD-OCT volumes in the training set. The weights of the CNN were randomly initialized using Xavier initialization and optimized using Adam optimization to minimize a binary cross-entropy loss. The CNN was trained for 10 epochs with a batch size of 250 and learning rate of 0.0001. Performance was evaluated on the hold-out validation set and the weights of the best-performing epoch were retained as the final weights of the CNN to be used during testing.

During testing, an *en face* probability map of EZ was generated by the trained CNN for each SD-OCT volume. The probability map was thresholded at 0.95 to obtain a binary map. A binary morphological operation was applied to fill any holes in the binary map and regions smaller than 0.025 mm^2 were removed to exclude small and discontinuous regions of EZ, to mimic the manual segmentation protocol. The hold-out validation set was used to establish the thresholds of 0.95 and 0.025 mm^2 indicated above. Figure 3 shows the automatic segmentation process.

Performance metrics

Several performance metrics were calculated to evaluate the performance of the automatic segmentation algorithm, using the manual segmentations as the gold standard. The Dice similarity coefficient (DSC) was calculated to measure the proportion of overlap between the manual and automatic *en face* segmentation maps. Measurements of the total EZ area (mm^2) and the maximum EZ length (μm) for each volume were also obtained from the segmentations. The absolute difference and Pearson's correlation between the manual and automatic measurements were calculated. A higher value indicates better performance for the DSC and Pearson's correlation, both of which range from 0 to 1, whereas a lower value indicates better performance for the absolute difference.

Generalizability of models trained on MacTel2

To investigate the generalizability of an existing model trained for a different application, we used the trained models from the MacTel2 study[22] to segment the EZ in *USH2A*-related RP. As there were six trained models from the MacTel2 study due to the six-fold cross-validation procedure, we used an ensemble of the six trained models by taking the mean of their predictions. As the models were trained to predict the probability of EZ defects in MacTel2, the probability map of EZ defects (p) was inverted ($1 - p$) to obtain the probability map of EZ in *USH2A*-related RP. The probability map was thresholded and postprocessed as described above, using the same thresholds of 0.95 and 0.025 mm^2.

Implementation

DOCTAD was implemented in Python with the TensorFlow[23] (Version 1.2.1) library. Statistical analysis was performed with MATLAB[24] (Version 9.5.0 R2018b).

RESULTS

Dataset

The dataset consisted of 127 SD-OCT volumes from 127 enrolled participants at the baseline visit. Participants in the RUSH2A study have been previously described.[25] All SD-OCT volumes consisted of 121 B-scans × 1536 A-scans within a 30° × 25° retinal area, with the exception of one volume with 768 B-scans, one volume with 49 B-scans × 1024 A-scans, and two volumes with 49 B-scans × 512 A-scans. All B-scans had a height of 496 pixels and an axial resolution of 3.87 μm/pixel. One participant was later considered ineligible because the EZ extended beyond the scan area, a study exclusion criterion for EZ analysis.

Quantitative analysis

Overall, there was good agreement between the manual and automatic segmentations with high DSC and Pearson's correlation, and low absolute differences. The average (mean ± SD) DSC was 0.79 ± 0.27 (median: 0.90). When the automatic measurements of total EZ area were compared to the manual measurements, the average absolute difference was 0.62 ± 1.41 (median: 0.22) mm^2 and the Pearson's correlation was 0.97. Similarly, when the automatic measurements of maximum EZ length were compared to

the manual measurements, the average absolute difference was 222 ± 288 (median: 126) μm and the Pearson's correlation was 0.97. Figure 4 shows the relationship between the measurements of the total EZ area and maximum EZ length obtained from the manual and automatic segmentations and the corresponding Pearson's correlation. Performance metrics are reported on all 126 eligible participants.

Qualitative analysis

Overall, there was good qualitative agreement between the manual and automatic segmentations. Figure 5 shows an example. The measurements of the total EZ area and EZ length were very similar and segmentation differences occurred only around the boundaries. This was true in general, despite the diverse disease manifestations and the presence of other abnormalities. Figure 6 shows various examples of segmentations in images with different disease manifestations, for which the EZ ranged from disrupted to clear and thin to thick, as well as the presence of other abnormalities such as intraretinal fluid, large central cysts, macular atrophy, adherent posterior hyaloid with vitreomacular traction, and epiretinal membrane with lamellar macular hole. Overall, the automatic segmentations were appropriate even in these extreme cases. Most segmentation differences occurred around the boundaries or in ambiguous regions where the EZ was disrupted and faintly visible but also not fully intact. As with any automatic segmentation algorithm, there were also a few cases of segmentation failures. Figure 7 shows examples where EZ was incorrectly identified (when clearly absent) or missed (when clearly present) in the automatic segmentations. However, such cases were very rare.

Generalizability of models trained on MacTel2

Using the trained models from the MacTel2 study, the average DSC was 0.43 ± 0.28 (median: 0.43). For measurements of the total EZ area, the average absolute difference was 1.85 ± 1.86 (median: 1.24) mm^2 and the Pearson's correlation was 0.89. For measurements of the maximum EZ length, the average absolute difference was 520 ± 528 (median: 395) μm and the Pearson's correlation was 0.85. We found that the predictions of the models trained on MacTel2 were considerably noisier when applied to *USH2A*-related RP. By applying an additional postprocessing step of simply removing all but the largest continuous region, which typically corresponded to the central EZ, the performance metrics improved. The average DSC was 0.45 ± 0.37 (median: 0.63). For measurements of the total EZ area, the average absolute difference was 1.34 ± 1.71 (median: 0.68) mm^2 and the Pearson's correlation was 0.95. For measurements of the maximum EZ length, the average absolute difference was 528 ± 430 (median: 409) μm and the Pearson's correlation was 0.91. Further improvements would likely be obtained with more complex modifications, which are beyond the scope of this paper. Figure 8 shows the relationship between the measurements of the total EZ area and maximum EZ length obtained from the manual and automatic segmentations and the corresponding Pearson's correlation. Overall, the performance of the models trained on MacTel2 was reasonable, but as expected, not as good as the performance of the models trained specifically for *USH2A*-related RP.

DISCUSSION

Many automatic algorithms have been developed to aid clinicians with medical image analysis. While an algorithm is often developed for and validated on a specific condition, explicit validation of the generalizability of the algorithm for application to other conditions is important for clinical applicability. In this article, we validated the generalizability of DOCTAD,[14, 22] a fully-automatic, deep learning-based, *en face* segmentation algorithm initially developed to segment EZ defects in MacTel2, for application to segment the presence of EZ in *USH2A*-related RP. To ensure that the performance evaluation would be clinically applicable, we used a diverse dataset from the large-scale, international, multi-center RUSH2A clinical study (NCT03146078).

Despite the diverse disease manifestations, DOCTAD segmented the EZ on SD-OCT volumes of participants with *USH2A*-related RP with good performance[26] without any need to modify the architecture or training procedure. Overall, DOCTAD achieved a high average DSC of 0.79 ± 0.27 (median: 0.90) compared to manual segmentations by trained Readers. This performance was comparable to the performance observed in the original MacTel2 studies, where DOCTAD achieved an average DSC of 0.79 ± 0.22 (median: 0.87), and which reproduced important clinical trial outcome measures.[14, 22] DOCTAD also achieved low absolute differences and high correlations of 0.97 for measurements obtained from the segmentations.

Even in extreme cases and in eyes with a variety of retinal abnormalities, severe segmentation failures were rare, demonstrating the robustness of the algorithm. Instead, segmentation differences occurred mostly around the boundaries of the EZ or in ambiguous regions where the EZ was disrupted and faintly visible but also not fully intact. These regions are difficult to segment even for expert Readers and often simply require a judgment call, which may differ between Readers. While such segmentation differences, due to the inherent bias or judgment in each segmentation method or Reader, may result in significant differences between the respective measurements at a single time point, measurements at a single time point are typically not the main clinical outcome measure. Instead, changes over time are more clinically relevant, as they reflect disease progression and treatment effects. Previous validation of DOCTAD's clinical applicability in a phase 2 clinical trial for MacTel2 showed that significant differences between manual and automatic measurements at a single time point were inconsequential, as each segmentation method was consistent over time and therefore able to accurately measure similar changes and treatment effects over time.[22] We will continue to further validate the performance of DOCTAD in the longitudinal natural history study of *USH2A*-related retinal degeneration over time as the RUSH2A clinical study progresses over four years. Analysis on a wider range of data and outcome measures will provide further insight into the reliability of the algorithm for clinical applications.

The performance of the models trained on MacTel2 was reasonable when applied to *USH2A*-related RP, as indicated by the high correlations of more than 0.90 for the measurements obtained from the segmentations. However, as indicated by the spread about the trendline, the differences between the manual and automatic segmentations were greater

for the models trained on MacTel2 (as shown in Figure 8), compared to the models trained specifically for *USH2A*-related RP (as shown in Figure 4). There are several possible reasons for these differences including different underlying disease pathology, different scanning protocols, and larger scanning field of view. However, perhaps the most important differences are the grading instructions and judgment criteria for manual segmentation in the different studies. Therefore, if the appropriate training dataset is available for a new application, training the algorithm for the specific application is highly recommended as there are often nuances in the manual segmentation approach that may not be reflected when using models trained for a different application. This approach would be straightforward with a generalizable methodology such as DOCTAD. However, in the absence of an appropriate training dataset, models trained for a different application may still be employed if the target biomarker is the same, although a certain degree of adjustment may be required to adhere to the specific segmentation protocols.

One limitation of this study is that we have only validated DOCTAD's generalizability to one other condition. However, DOCTAD also has the potential to be used to segment other biomarkers of interest in other conditions, such as geographic atrophy or drusen in age-related macular degeneration. We will continue to validate DOCTAD's generalizability to these other conditions in our future work.

In conclusion, we validated the generalizability of DOCTAD, a fully-automatic, deep learning-based, *en face* segmentation algorithm, that can be applied to retinal diseases other than MacTel2, the disease on which it was originally developed for. We showed that DOCTAD could be applied to segment the EZ in *USH2A*-related retinal degeneration with good performance on a diverse dataset from a large-scale clinical study. We also recommend that algorithms from other investigators should be validated for different aspects of clinical applicability, including generalizability to other conditions, to increase their potential for real-world clinical use.

ACKNOWLEDGEMENTS

The source of the data is the Foundation Fighting Blindness Consortium, but the analyses, content and conclusions presented herein are solely the responsibility of the authors and may not reflect the views of the Foundation Fighting Blindness.

Funding:

National Institutes of Health (P30 EY005722). Research to Prevent Blindness (Unrestricted Grant to Duke University). Funding to support the Rate of Progression of *USH2A*-related Retinal Degeneration was provided by the Foundation Fighting Blindness Consortium.

REFERENCES

1. Mathur P, Yang J. Usher syndrome: hearing loss, retinal degeneration and associated abnormalities. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease 2015;1852(3):406–420. [PubMed: 25481835]
2. Eudy JD, Weston MD, Yao S, et al. Mutation of a gene encoding a protein with extracellular matrix motifs in Usher syndrome type IIa. Science 1998;280(5370):1753–1757. [PubMed: 9624053]

3. Pierrache LH, Hartel BP, Van Wijk E, et al. Visual prognosis in USH2A-associated retinitis pigmentosa is worse for patients with usher syndrome type IIa than for those with nonsyndromic retinitis pigmentosa. Ophthalmology 2016;123(5):1151–1160. [PubMed: 26927203]
4. Fuster-García C, García-García G, González-Romero E, et al. USH2A gene editing using the CRISPR system. Molecular Therapy-Nucleic Acids 2017;8:529–541. [PubMed: 28918053]
5. Jacobson SG, Cideciyan AV, Aleman TS, et al. Usher syndromes due to MYO7A, PCDH15, USH2A or GPR98 mutations share retinal disease mechanism. Human molecular genetics 2008;17(15):2405–2415. [PubMed: 18463160]
6. Mukherjee D, Lad EM, Vann RR, et al. Correlation between macular integrity assessment and optical coherence tomography imaging of ellipsoid zone in macular telangiectasia type 2. Invest. Ophth. Vis. Sci. 2017;58(6):291–299.
7. Chew EY, Clemons TE, Jaffe GJ, et al. Effect of ciliary neurotrophic factor on retinal neurodegeneration in patients with macular telangiectasia type 2: a randomized clinical trial. Ophthalmology 2019;126(4):540–549. [PubMed: 30292541]
8. de Sisternes L, Hu J, Rubin DL, Leng T. Visual prognosis of eyes recovering from macular hole surgery through automated quantitative analysis of spectral-domain optical coherence tomography (SD-OCT) scans. Invest. Ophth. Vis. Sci. 2015;56(8):4631–4643.
9. Itoh Y, Vasanji A, Ehlers JP. Volumetric ellipsoid zone mapping for enhanced visualisation of outer retinal integrity with optical coherence tomography. Brit. J. Ophthalmol. 2016;100(3):295–299. [PubMed: 26201354]
10. Zhu W, Chen H, Zhao H, et al. Automatic three-dimensional detection of photoreceptor ellipsoid zone disruption caused by trauma in the OCT. Sci. Rep. 2016;6:25433. [PubMed: 27157473]
11. Wang Z, Camino A, Zhang M, et al. Automated detection of photoreceptor disruption in mild diabetic retinopathy on volumetric optical coherence tomography. Biomed. Opt. Express 2017;8(12):5384–5398. [PubMed: 29296475]
12. Banaee T, Singh RP, Champ K, et al. Ellipsoid zone mapping parameters in retinal venous occlusive disease with associated macular edema. Ophthalmol. Retina 2018.
13. Camino A, Wang Z, Wang J, et al. Deep learning for the segmentation of preserved photoreceptors on en face optical coherence tomography in two inherited retinal diseases. Biomed. Opt. Express 2018;9(7):3092–3105. [PubMed: 29984085]
14. Loo J, Fang L, Cunefare D, et al. Deep longitudinal transfer learning-based automatic segmentation of photoreceptor ellipsoid zone defects on optical coherence tomography images of macular telangiectasia type 2. Biomed. Opt. Express 2018;9(6):2681–2698. [PubMed: 30258683]
15. Lang A, Carass A, Bittner AK, et al. Improving graph-based OCT segmentation for severe pathology in Retinitis Pigmentosa patients. Medical Imaging 2017: Biomedical Applications in Molecular, Structural, and Functional Imaging. 2017.
16. Liu Y, Carass A, He Y, et al. Layer boundary evolution method for macular OCT layer segmentation. Biomed. Opt. Express 2019;10(3):1064–1080. [PubMed: 30891330]
17. He Y, Carass A, Liu Y, et al. Structured layer surface segmentation for retina OCT using fully convolutional regression networks. Med. Image. Anal. 2021;68:101856. [PubMed: 33260113]
18. Yang Q, Reisman CA, Chan K, et al. Automated segmentation of outer retinal layers in macular OCT images of patients with retinitis pigmentosa. Biomed. Opt. Express 2011;2(9):2493–2503. [PubMed: 21991543]
19. Wang Y-Z, Cao A, Birch DG. Evaluation of a UNet Convolutional Neural Network (CNN) for Automatic Measurements of Ellipsoid Zone (EZ) Area and Photoreceptor Outer Segment (POS) Volume in X-Linked Retinitis Pigmentosa (xlRP). Invest. Ophth. Vis. Sci. 2021;62(8):2134–2134.
20. De Silva T, Jayakar G, Grisso P, et al. Deep-learning based automatic detection of ellipsoid zone loss in SD-OCT for hydroxychloroquine retinal toxicity screening. Ophthalmology Science 2021:100060.
21. Loo J, Cai CX, Choong J, et al. Deep learning-based classification and segmentation of retinal cavitations on optical coherence tomography images of macular telangiectasia type 2. Brit. J. Ophthalmol. 2020.

22. Loo J, Clemons TE, Chew EY, et al. Beyond Performance Metrics: Automatic Deep Learning Retinal OCT Analysis Reproduces Clinical Trial Outcome. Ophthalmology 2020;127(6):793–801. [PubMed: 32019699]

23. Abadi M, Barham P, Chen J, et al. TensorFlow: A System for Large-Scale Machine Learning. OSDI. 2016.

24. MATLAB. version 9.5.0 (R2018b). The MathWorks Inc. 2018.

25. Duncan JL, Liang W, Maguire MG, et al. Baseline visual field findings in the RUSH2A study: associated factors and correlation with other measures of disease severity. Am. J. Ophthalmol. 2020;219:87–100. [PubMed: 32446738]

26. Mukaka MM. A guide to appropriate use of correlation coefficient in medical research. Malawi Medical Journal 2012;24(3):69–71. [PubMed: 23638278]

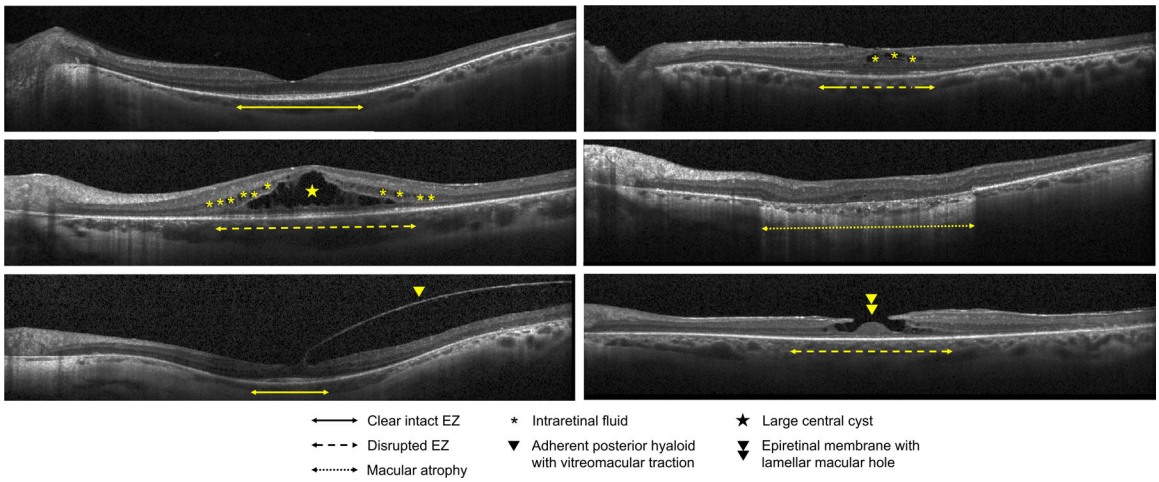

Figure 1.

Example images from the RUSH2A clinical study that exhibited a wide range of disease severity and macular findings. This diversity increases the segmentation complexity and would provide a more robust evaluation of the algorithm's performance.

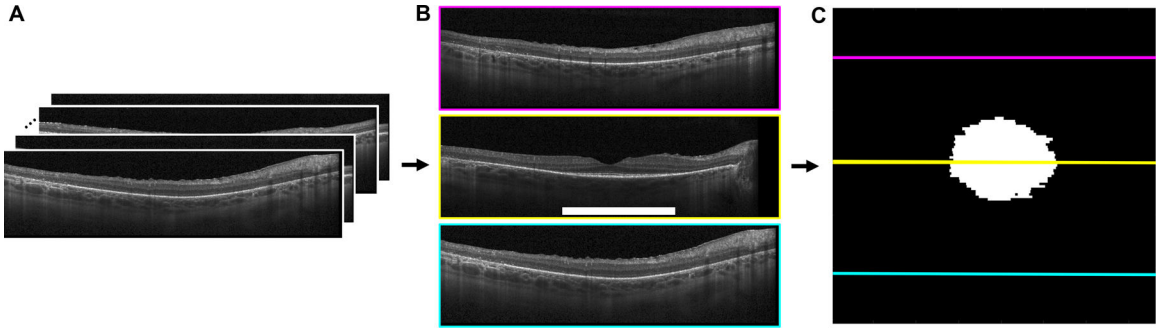

Figure 2.

Illustration of the manual segmentation process. **A:** SD-OCT volume. **B:** Examples of B-scans with the segmentation of EZ (white). **C:** *En face* segmentation map of EZ. The colored lines correspond to the position of the B-scan outlined in the same color.

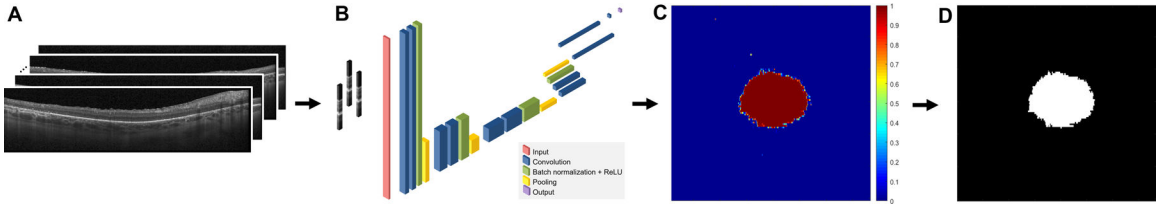

Figure 3.

Illustration of the automatic segmentation process. **A:** SD-OCT volume. **B:** CNN trained to classify clusters of A-scans with the absence or presence of EZ. **C:** *En face* probability map of EZ generated by the trained CNN. **D:** *En face* segmentation map of EZ obtained via thresholding and post-processing.

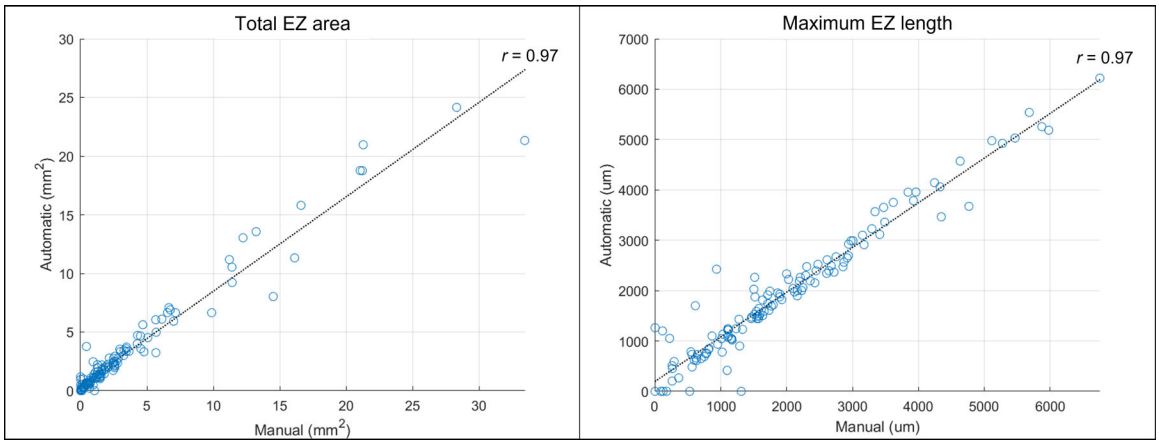

Figure 4.

Scatterplots showing the relationship between the measurements of the total EZ area and maximum EZ length obtained from the manual and automatic segmentations and the corresponding Pearson's correlation, *r* on all 126 eligible participants with *USH2A*-related RP.

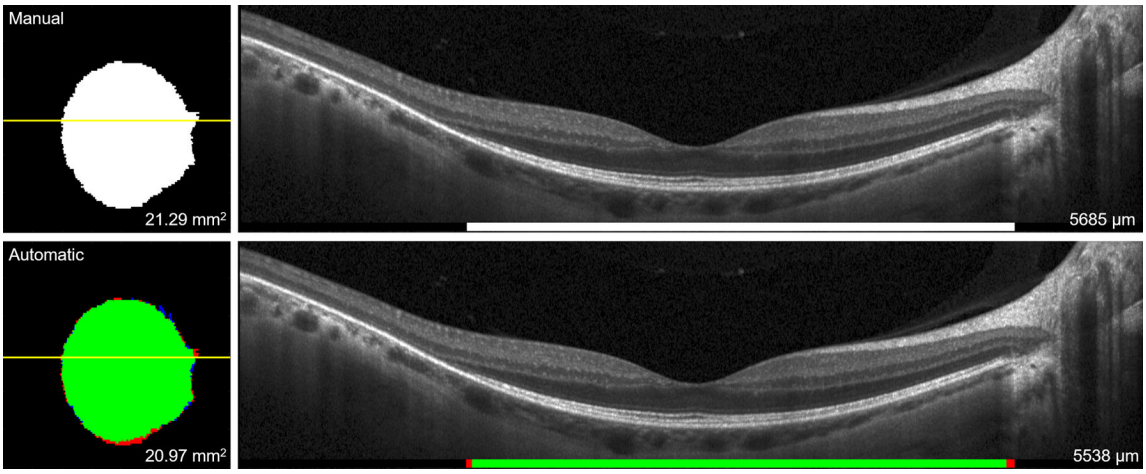

Figure 5.

Example of good agreement between the manual and automatic segmentations. **Left:** *En face* segmentation maps of EZ. The measurements of the total EZ area are shown in the bottom corner. **Right:** Segmentations of EZ in the B-scan corresponding to the yellow line. The measurements of the EZ length are shown in the bottom corner. The colors in the automatic segmentations correspond to the true positives (green), false positives (blue), and false negatives (red).

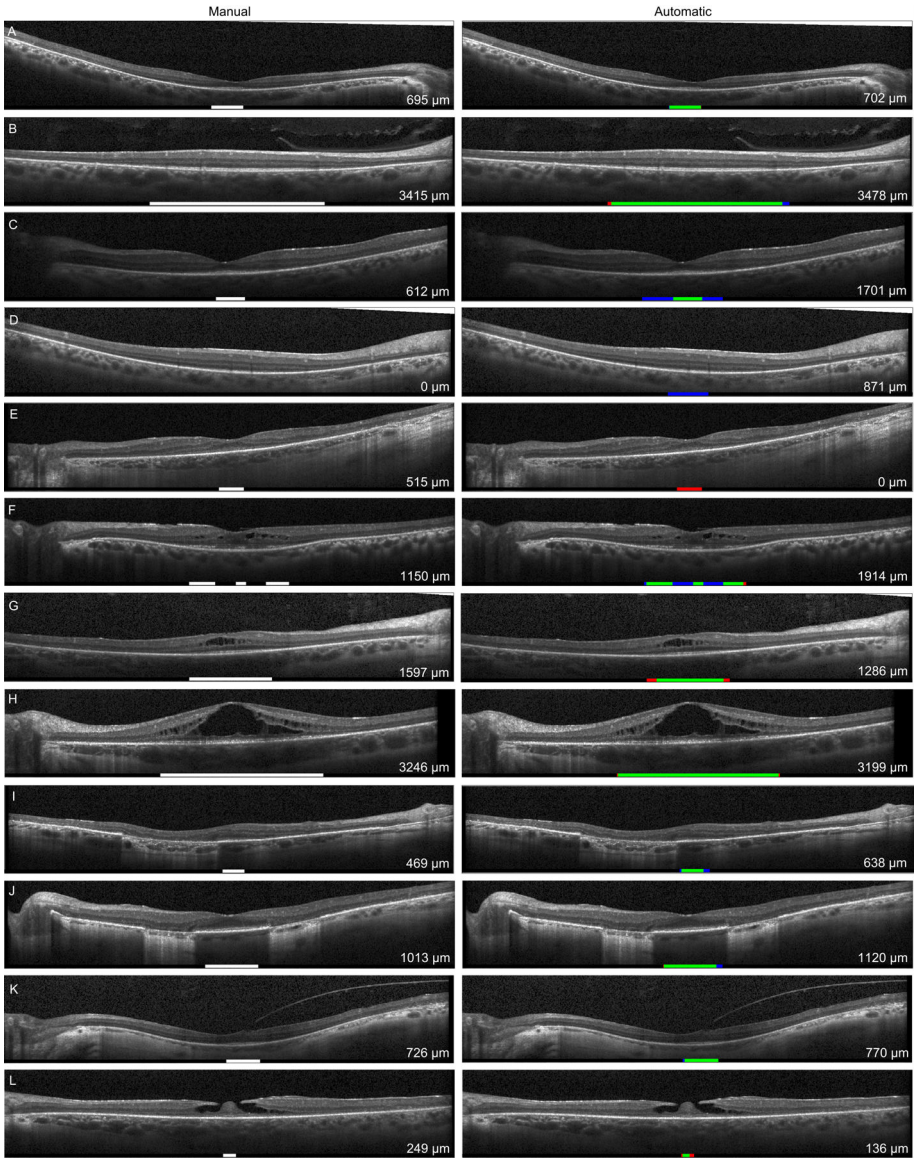

Figure 6.

Examples of segmentations in images with different disease manifestations for which the EZ ranged from disrupted to clear and thin to thick (A – E), as well as the presence of other abnormalities such as intraretinal fluid (F – H), large central cyst (H), macular atrophy (I, J), adherent posterior hyaloid with vitreomacular traction (K), and epiretinal membrane with lamellar macular hole (L). The measurements of the EZ length are shown in the bottom corner. The colors in the automatic segmentations correspond to the true positives (green), false positives (blue), and false negatives (red).

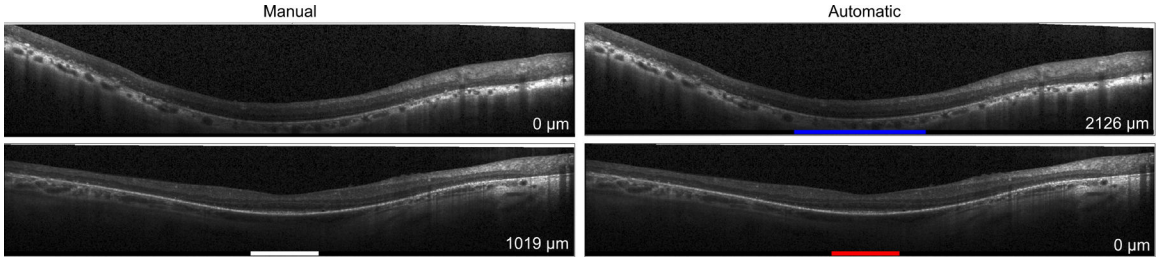

Figure 7.

Examples of rare segmentation failures where EZ was incorrectly identified (top) or missed (bottom) in the automatic segmentations. The measurements of the EZ length are shown in the bottom corner. The colors in the automatic segmentations correspond to the false positives (blue) and false negatives (red).

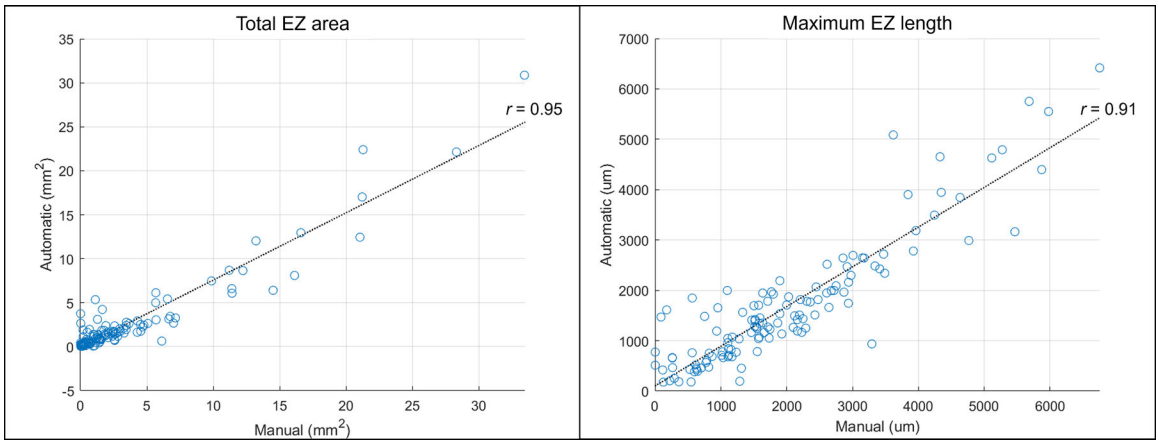

Figure 8.

Scatterplots showing the relationship between the measurements of the total EZ area and maximum EZ length obtained from the manual and automatic segmentations and the corresponding Pearson's correlation, *r* on all 126 eligible participants with *USH2A*-related RP. The automatic segmentations were obtained using the models trained on images from participants with MacTel2.