Deep Learning-Based Automatic Detection of Ellipsoid Zone Loss in Spectral-Domain OCT for Hydroxychloroquine Retinal Toxicity Screening

Tharindu De Silva, PhD,1 Gopal Jayakar, BS,1 Peyton Grisso, BS,1 Nathan Hotaling, PhD,1,2 Emily Y. Chew, MD,1 Catherine A. Cukras, MD, PhD1

Purpose: Retinal toxicity resulting from hydroxychloroquine use manifests photoreceptor loss and disruption of the ellipsoid zone (EZ) reflectivity band detectable on spectral-domain (SD) OCT imaging. This study investigated whether an automatic deep learning-based algorithm can detect and quantitate EZ loss on SD OCT images with an accuracy comparable with that of human annotations.

Design: Retrospective analysis of data acquired in a prospective, single-center, case-control study.

Participants: Eighty-five patients (168 eyes) who were long-term hydroxychloroquine users (average exposure time, 14 ± 7.2 years).

Methods: A mask region-based convolutional neural network (M-RCNN) was implemented and trained on individual OCT B-scans. Scan-by-scan detections were aggregated to produce an en face map of EZ loss per 3-dimensional SD OCT volume image. To improve the accuracy and robustness of the EZ loss map, a dual network architecture was proposed that learns to detect EZ loss in parallel using horizontal (horizontal mask region-based convolutional neural network [M-RCNNH]) and vertical (vertical mask region-based convolutional neural network [M-RCNNV]) B-scans independently. To quantify accuracy, 10-fold cross-validation was performed.

Main Outcome Measures: Precision, recall, intersection over union (IOU), F1-score metrics, and measured total EZ loss area were compared against human grader annotations and with the determination of toxicity based on the recommended screening guidelines.

Results: The combined projection network demonstrated the best overall performance: precision, 0.90 ± 0.09; recall, 0.88 ± 0.08; and F1 score, 0.89 ± 0.07. The combined model performed superiorly to the M-RCNNH only model (precision, 0.79 ± 0.17; recall, 0.96 ± 0.04; IOU, 0.78 ± 0.15; and F1 score, 0.86 ± 0.12) and M-RCNNV only model (precision, 0.71 ± 0.21; recall, 0.94 ± 0.06; IOU, 0.69 ± 0.21; and F1 score, 0.79 ± 0.16). The accuracy was comparable with the variability of human experts: precision, 0.85 ± 0.09; recall, 0.98 ± 0.01; IOU, 0.82 ± 0.12; and F1 score, 0.91 ± 0.06. Automatically generated en face EZ loss maps provide quantitative SD OCT metrics for accurate toxicity determination combined with other functional testing.

Conclusions: The algorithm can provide a fast, objective, automatic method for measuring areas with EZ loss and can serve as a quantitative assistance tool to screen patients for the presence and extent of toxicity. Ophthalmology Science 2021;1:100060 Published by Elsevier on behalf of the American Academy of Ophthalmology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Supplemental material available at www.ophthalmologyscience.org.

Hydroxychloroquine is a widely used drug to treat autoimmune diseases such as systemic lupus erythematosus, Sjögren’s syndrome, and rheumatoid arthritis. One of the major known side-effects among long-term users of the drug is retinal toxicity that can result in permanent damage to photoreceptors and retinal pigment epithelium, eventually leading to irrecoverable central vision loss. Although this side-effect is estimated to occur in 7.5% of patients taking the drug for more than 10 years,1 currently, no treatment exists, and the damage tends to continue even after the cessation of the drug,2,5 making screening essential. The American Academy of Ophthalmology (AAO) recommends 2 main screening methods, spectral-domain (SD) OCT imaging and functional tests such as visual field testing, with the goal of recognizing early definitive signs of toxicity to prevent vision loss.7 Spectral-domain OCT imaging allows for the evaluation of the retinal layers including the outer retina and plays a vital role in screening for evidence of the structural changes induced by drug toxicity. Detection of the ellipsoid zone (EZ) band and any associated EZ loss have been proposed as outcome measures of disease progression in several degenerative diseases5,9 because these indicate the deterioration of
photoreceptors. Ellipsoid zone loss area on en face maps of eyes with hydroxychloroquine toxicity also has been shown to correlate with mean deviation of visual function.

Although SD OCT is considered to be an objective method depicting structural changes, the current clinical hydroxychloroquine screening uses qualitative inspection of individual OCT B-scans to identify areas of localized photoreceptor thinning and EZ loss (interruption or discontinuity of the EZ), which represent and denote definitive evidence of retinal toxicity. Although classical presentation of apparent toxicity is described as bilateral bull’s-eye maculopathy, where the retinal layers have degenerated in the shape of a parafoveal ring sparing a foveal island, early-stage disruptions in EZ can be subtle and miniscule. Variable degrees of severity present on OCT images, suboptimal image quality, and interpretation of images by those with different levels of clinical expertise (screening performed by retinal specialists, ophthalmologists, and, in some settings, optometrists) can induce subjectivity, variability, and error to current diagnostics. A fully automatic method to detect and quantify loss of the EZ band from SD OCT imaging thus would add agnostics. A fully automatic method to detect and quantify EZ loss accurately in this data set, which includes patients who did not exhibit any signs of toxicity, those with subtle, mild cases of toxicity with minimal functional loss, and those with severe toxicity with significant functional loss. We examined the ability of automatically derived quantitative metrics to facilitate the screening process by making toxicity determinations in an objective manner. Finally, we explored the relationship of the measured EZ loss area and functional measurements such as visual field mean deviation and visual acuity, where EZ loss maps could provide useful insights into the visual function deficiencies arising from toxicity.

Methods

Participants

Data were collected as part of a National Institutes of Health institutional review board–approved clinical study (clinicaltrials.gov identifier, NCT01145196) from patients who were long-term (>5 years) users of hydroxychloroquine. Written informed consent was obtained from all participants and the study protocols adhered to the tenets of the Declaration of Helsinki and the Health Insurance Portability and Accountability Act. Eighty-five participants with a mean age of 59 ± 12 years, 93% of whom were women, who were exposed to hydroxychloroquine for a mean of 14 ± 7.2 years were included in the analysis. Patients underwent a comprehensive ocular examination including AAO 2016 revised screening tests: multimodal imaging, multifocal electroretinography (mERG) testing, and perimetry testing. During study visits, SD OCT images were acquired using the Heidelberg Spectralis HRA+OCT system (Heidelberg Engineering, Inc) following 2 scan protocols. The first consisted of a single horizontal 30° scan (768 × 496) through the fovea with 100 frame averaging. The second was a volumetric macular cube scan (voxel dimensions, 768 × 496 × 121; voxel spacing, 11.9 × 3.9 × 64.0 μm) spanning 30° horizontally and 25° vertically. The images exhibited variable image quality representative of practical clinical settings acquired during a period of 5 or more years. The signal-to-noise ratio (SNR) as measured by the vendor device exhibited a range of 17.8 to 31.1 dB with mean ± standard deviation SNR of 26.7 ± 2.7 dB. The scanning protocol was set with a frame averaging of 2 in 98.5% percent of the B-scans. The remainder had frame averaging ranging
from 5 to 25. The study data set included both eyes of each participant, except for 2 eyes that were excluded because of unavailable or suboptimal-quality OCT scans, leaving 168 eyes available for analysis.

**Determination of Toxicity**

Toxicity was determined based on the AAO recommendations, including the findings from a combination of objective (SD OCT or mERG) and subjective (Humphrey visual field) testing. Evidence for toxicity in visual field tests was defined as having either 3 contiguous abnormal points on the pattern deviation map or a full-ring scotoma.26 Horizontal SD OCT foveal B-scans were inspected for evidence of EZ loss. In mERG, reduced central amplitude (< 35 nV2) or abnormal ring ratio (> 2.6) were used as criteria for evidence of toxicity.27,28 Fifty-five patients (110 eyes) were classified as unaffected (i.e., without toxicity), whereas 30 patients (58 eyes) were identified as affected and having toxicity. For details on toxicity determinations, see Supplemental Table 1.

**Automatic Algorithm for Ellipsoid Zone Loss Detection**

Forty-two eyes (21 patients) of the 30 affected participants exhibited significant EZ loss, defined as 100 μm or more of loss in the foveal SD OCT scan. Two trained human experts (G.J., P.G.), with verification by a retina specialist (C.C.), manually annotated regions of EZ loss in the 5082 horizontal scans of the 21 eyes, producing 3477 B-scans with contours of EZ loss and 1605 horizontal scans without evidence of EZ loss. Ellipsoid zone loss was defined as complete absence of the EZ band, such that areas of EZ band attenuation without complete absence (as occurs in some cases of early severity, adjacent to areas of frank disruption, or as occurs in shadowing under large retina vessels) did not qualify. In addition, 33,432 vertical scans were interpolated from the volume scans, and ground truth vertical annotations were derived automatically via interpolation of manually defined horizontal annotations. The human expert annotation of regions with EZ loss is considered the gold standard for the automatic algorithm training and evaluation, and henceforth is referred to as ground truth. These ground truth annotations were provided as input to the deep learning algorithm.

We devised a 2-step strategy to generate an en face EZ loss map for each volumetric SD OCT scan. The first step consisted of an algorithm for scan-by-scan EZ loss detection independently in each B-scan from the SD OCT volume. We implemented a mask region-based convolutional neural network (M-RCNN)29,30 to predict the EZ loss using manually annotated SD OCT B-scans capturing EZ loss, as shown in Figure 1. ResNet-50 was used as the backbone for feature selection with 256 hidden layers. Pretrained weights (trained on the Common Objects in Context (COCO) data set with natural images) was used to initialize the model. The network was trained by optimizing for multitask losses for box detection, classification, and mask labelling. To provide ground truth for the detection task, boxes were generated encapsulating each manually defined EZ loss segment. The encapsulating boxes had a range of sizes and aspect ratios, depending on the extent of EZ loss and the slope of EZ layer within the B-scan. The classification in this instance was a binary task where the presence of EZ loss versus no loss was determined for each box proposal. Stochastic gradient descent optimizer was used with a learning rate of 0.0005, momentum of 0.9, and weight decay of 0.0005. These hyperparameters, including scale (5 values between 32 and 512) and aspect ratios (0.5, 1, and 2), were set empirically without using an explicit validation set. In each fold, training was performed for 20 epochs with random generation of training data samples until the error measured in the training set converged. The network was implemented in PyTorch version 1.6. For an input B-scan, the network predicted and annotated the corresponding region of EZ loss, as shown in Figure 1.

In the second step, scan-by-scan detections were projected onto an en face image to generate a 2D map representing the regions of EZ loss in the 3D OCT volume. To improve the accuracy and robustness of estimating the en face EZ loss map, we devised 2-fold additions to the overall network architecture, as shown in Figure 2. The first addition was by way of redundancy, where a dual-path network31 generated 2 EZ loss maps using horizontal and vertical scans extracted from the same 3D OCT volume training 2 independent networks in parallel in horizontal (horizontal mask region-based convolutional neural network [M-RCNNH]) and vertical (vertical mask region-based convolutional neural network [M-RCNNV]) directions, as shown in Figure 2. The second add-on to the network is an aggregation network that operates on horizontally and vertically derived dual-projection maps combining the outputs of M-RCNNH and M-RCNNV, mitigating any inconsistencies arising from independent scan-by-scan detections. The output of this combined projection network (CPN) made the final consensus estimate for the presence of EZ loss at each location in the en face map. This approach could be more robust because it can identify any spatially inconsistent patterns of spurious detections across parallel scan-by-scan detections. The CPN was a custom network implementation of 3 convolution layers each matching the dimensions of the EZ loss map (768 × 768 pixels) without any downsampling or upsampling. Relu and MaxPool operations were applied at the end of the first 2 convolutional layers. The output of the third convolutional layer was derived directly as the EZ loss map. This provided a regression output, and a thresholding operation was used to obtain the final binary EZ loss map. This network was trained with mean squared error loss using stochastic gradient descent optimizer with a learning rate of 0.001 and momentum of 0.9. The training was performed for 10 fixed number of epochs until the training error converged.

**Experiments**

The deep learning model was developed using images from 42 eyes, with 90% of the data used for training while testing with the remaining 10%. The model with 2 steps was trained serially, where the first stage of the network was trained using individual B-scans and the second stage of the network was trained after freezing all the weights of the first stage of the network and using all the B-scans from a volume. In both stages, individual scans from the same patient were not included in both training and test sets simultaneously to mitigate possible bias during learning. Ten-fold cross-validation was performed and the accuracy of detecting EZ loss in en face 2D maps was quantified using precision (positive predictive value) (TP/(TP+FP), recall (sensitivity) (TP/(TP+FN)), intersection over union (IOU) (TP/(TP+FP+FN)), and F1 score (Dice) (2(精度+召回)/(精度×召回)) metrics. To compare performance with a retinal layer segmentation-based method, a previously published deep learning model was used to generate pixel labels for retinal layers. The model was trained using labelled data from patients with age-related macular degeneration, and the layer thickness was set to 0 in regions with loss to compute EZ loss accuracy. The model Deep Lab v3 (DLabv3) comprised a convolutional neural network Deeplabv3 with a ResNet50 backbone (the same feature selection backbone used in the proposed model) and operated independently on individual B-scans. All pixels were labelled into 7 different classes (inner retina, outer nuclear layer, inner segments,
outer segments, retina pigment epithelium drusenoid complex (RPEDC), choroid, and background), as described in Pfau et al.\textsuperscript{27} The presence of the boundary between inner and outer segments was defined as EZ presence, and regions without this boundary were identified as EZ loss. En face EZ loss maps were generated by projecting the presence or absence of EZ loss along each A-scan.

Annotating EZ loss by human experts could be subjective in some scenarios where the EZ layer is deteriorated, but not completely lost. This variability was measured in a separate experiment by repeatedly annotating a subset of 7 OCT volumes (847 B-scans) by the 2 human experts independently to provide quantification for the human grader variation.

After training and testing the model on the cohort of 21 eyes with definitive signs of toxicity, we generated en face EZ loss maps for all the participants in the study, including unaffected eyes and the eyes with the mild findings (<100-μm EZ loss in foveal slice). The verification of the absence of EZ loss on en face maps of unaffected patients not included in the development of the algorithm would test its validity to operate on the entire patient population, most of whom do not show any sign of EZ loss. We then evaluated how a quantitative metric derived from the EZ loss map

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{Scan-by-scan ellipsoid zone (EZ) loss detection and segmentation using a mask region-based convolutional neural network (RCNN). SD = spectral-domain. A, Original SD-OCT B-scan; B, Human annotated ground truth locations corresponding to EZ loss; C, Mask-RCNN network predicting the EZ loss regions with SD-OCT B-scan as input.}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure2.png}
\caption{Diagram showing combined projection network predicting en face ellipsoid zone (EZ) loss map by aggregating scan-by-scan detections in a horizontal mask region-based convolutional neural network (M-RCNN\textsubscript{h}) and vertical mask region-based convolutional neural network (M-RCNN\textsubscript{v}).}
\end{figure}
(i.e., total EZ loss area measured within a 6-mm diameter from fovea) correlated with other functional test results.

Results

Accuracy Validation of En Face Ellipsoid Zone Loss Map Generation

Figure 3 shows en face EZ loss maps generated from different models with a range of EZ loss areas. In slice-by-slice detections in horizontal and vertical networks, the EZ loss map contained nonsmooth regions with horizontal and vertical streaking artifacts resulting from the lack of spatial consistency between slices. Ground truth projections also exhibited this artifact because the manually annotated EZ loss end points did not line up across adjacent B-scans perfectly. Overall, the combined projection network predicted output maps better resembling the ground truth, successfully mitigating some spurious detections inconsistent across horizontal and vertical maps. The streaking artifacts in horizontal and vertical networks also were mitigated, producing a more regularized output for EZ loss regions.

Figure 4 compares the performance of the deep learning models evaluated in this study with the ground truth provided by manual contouring. The DLab3 model exhibited performance with the following mean ± standard deviation values: precision, 0.79 ± 0.21; recall, 0.72 ± 0.23; IOU, 0.64 ± 0.25; and F1 score, 0.74 ± 0.23. This model was trained in patients with different levels of age-related macular degeneration severity that include instances of EZ layer deterioration and indicated the ability to detect EZ loss successfully in patients taking hydroxychloroquine. This serves as a baseline for comparing the performance of the proposed detection-based models. The M-RCNNH achieved precision of 0.79 ± 0.17, recall of 0.96 ± 0.04, IOU of 0.78 ± 0.15, and F1 score of 0.86 ± 0.12. The M-RCNNV demonstrated slightly inferior performance, with precision of 0.71 ± 0.21, recall of 0.94 ± 0.06, IOU of 0.69 ± 0.21, and F1 score of 0.79 ± 0.16. This decrease in performance of the M-RCNNV can be attributed to the suboptimal image quality of synthetic (i.e., reconstructed) vertical B-scans extracted from horizontally acquired OCT volumes. Average precision for the box detection task at 0.5 IOU in the test set was 49.6% and 33.5% for the M-RCNNH and M-RCNNV networks, respectively. The combined model, CPN, demonstrated significantly better performance than either the M-RCNNH or M-RCNNV networks, with precision of 0.90 ± 0.09, recall of 0.88 ± 0.08, IOU of 0.82 ± 0.12, and F1 score of 0.89 ± 0.07 (P < 0.001 vs. M-RCNNH, paired t test), confirming the hypothesis that combining horizontal and vertical detection improves the robustness of en face EZ loss estimation. Overall, the models yielded superior recall compared with precision, indicating a smaller false-negative rate compared with the false-positive rate.

The data set contained a cohort of patients with a range of toxicities, and therefore a variable range of EZ loss, with
some moderate toxicity eyes exhibiting small areas of EZ loss in a partial arc around the fovea (Fig 3, top row), and some others at late stages with EZ deterioration visible in a large circular area of EZ loss encompassing most of the macula (Fig 3, middle and bottom rows). According to manual annotations, the mean ± standard deviation area of EZ loss measured within a 6-mm diameter from the fovea in the images was 16.6 ± 9.1 mm² and ranged between 0.1 and 30.2 mm². The accuracy in predicting the area of EZ loss in the image using the combined projection

![Image](image-url)
network was $1.1 \pm 1.5 \text{ mm}^2$. Figure 5A shows the comparison of human expert EZ loss against the algorithm-estimated EZ loss with excellent correlation ($R^2 = 0.98$), with error in estimating loss remaining fairly unchanged as a function of EZ loss area. A Bland-Altman plot (Fig 5B) indicated a slight overestimation of the EZ loss area, with 95% limits of agreement ranging between $-3.1$ and $2.9 \text{ mm}^2$. Thus, the network could predict EZ deterioration accurately in a variable range, which could be useful in classifying the disease severity of the patients and integrating to the clinical decision-making process.

Although the ground truth of manual contouring is the existing gold standard for EZ loss determination, variability exists among human experts in annotating the precise borders of the regions of EZ loss. A subset of 7 OCT cube scans independently graded by 2 graders (G.J., P.G.) allowed the quantification of variability between graders, which produced precision of $0.85 \pm 0.09$, recall of $0.98 \pm 0.01$, IOU of $0.82 \pm 0.12$, and F1 score of $0.91 \pm 0.06$ ($P = 0.80$, paired t test, indicating no statistically significant difference). In this subset, the measured areas between graders showed error of $1.1 \pm 0.8 \text{ mm}^2$, and the 1-way analysis of variance failed to find the statistical significance ($P = 0.95$) among the area measurements computed with 2 expert grader annotations and the algorithm, as shown in Figure 6B. The variability of human expert annotations is a fundamental limitation to the optimal accuracy attainable with a learning-based model and contextualizes the performance of CPN in the experiments that approached the variability of human experts.

**En Face Ellipsoid Zone Loss Maps for Clinical Screening**

Although the model was developed using a subset of patients in the study with significant EZ loss (eyes with $>100\mu$m EZ loss on the foveal SD OCT slice), the entire set of eyes was used to assess the usefulness of the automatic EZ loss detection algorithm in assisting clinical screening for toxicity. For all study participants, toxicity was determined based on a combination of evidence as demonstrated in 1 objective test (SD OCT or mfERG criteria) and 1 subjective test (visual field). Based on the AAO-recommended clinical determination of toxicity, 58 of the 158 eyes were affected, whereas 110 eyes were classified as unaffected. Figure 7 compares the total EZ loss area distributions measured in affected and unaffected groups. The algorithm did not identify substantial EZ loss regions (mean ± standard deviation EZ loss area, $0.01 \pm 0.07 \text{ mm}^2$) in the 110 unaffected eyes, validating our model in successfully confirming the absence of EZ loss in this group of patients. Inspection of the 8 of 110 eyes that exhibited nonzero EZ loss area ($0.1-0.5 \text{ mm}^2$) revealed regions with fading of the EZ layer without complete loss arising from shadowing or peripapillary atrophy. The algorithm detected clear signs of EZ loss areas (mean±standard deviation, $15.71 \pm 9.49 \text{ mm}^2$) in eyes with toxicity. Sixteen of the 58 eyes with toxicity showed minimal (range, $0-<100 \text{ mm}$) EZ loss in the foveal B scan and were not part of the model development. However, in these 16 eyes, the model did detect evidence of EZ loss in 7 eyes (range, $0.1-3.9 \text{ mm}^2$) and did not detect evidence of EZ loss in the remaining 9 eyes.

The area under the receiver operating characteristic curve of differentiating between unaffected and affected groups (determined using AAO screening recommendations) using the EZ loss area as a metric was found to be 0.91. With an optimal threshold at EZ area of 0.007 mm$^2$, this corresponded to a classification accuracy of 89.9% (151/168 eyes) with a false-positive rate of 7.3% (8/110 eyes) and a false-negative rate of 15.5% (9/58 eyes). All the false-negative findings were triggered because of eyes that were determined clinically to have toxicity, but for which the algorithm did not detect EZ loss. The human grader qualitative inspection also revealed no apparent EZ loss in these images, confirming the accuracy of the algorithm. These
patients were identified as affected because of the positive visual field findings and mfERG results, demonstrating the complexities of toxicity determination in certain cases. Overall, the algorithm detected all cases of definitive toxicity with EZ disruption, even those with very small areas of EZ loss. Thus, regions of EZ loss detected by the algorithm combined with other functional testing provide an automatic, fast, robust method to improve the efficacy of clinical screening.

**Ellipsoid Zone Loss Correlations with Visual Function**

In addition to facilitating clinical screening for toxicity, automatically generated EZ loss maps could help to investigate the relationship between structure and function. To illustrate such potential applications of the algorithm, we analyzed the relationship between total EZ loss area and the mean deviation on the 10-2 Humphrey visual field. Figure 8A demonstrates a strong negative correlation ($R^2 = -0.81$) in the relationship between EZ loss area (measured using the algorithm) and Humphrey visual field mean deviation, indicating worsening function with larger EZ loss. Additionally, analysis demonstrated that the visual acuity is impacted when the EZ loss occurs close (within 0.2 mm) to the fovea (Fig 8B). Thus, automatic tools for EZ loss generation provide a useful method for the exploration of structural changes underpinning visual function, and the usefulness would be even greater in large-scale studies where manual annotation and inspection are not feasible.

**Discussion**

This work reports an automatic learning-based model to estimate EZ loss in patients at risk of retinal toxicity arising from long-term hydroxychloroquine use. The M-RCNN with transfer learning from natural images successfully identified regions of EZ loss in individual B-scans. Scan-by-scan detections were aggregated to construct an EZ loss map representing the complete loss for the eye. The accuracy and robustness of the EZ loss map were improved by implementing a dual architecture that operates in a redundant manner on horizontal and vertical B-scans from the same image and then aggregates the dual EZ loss maps with an additional set of convolution layers. Although both horizontal and vertical B-scans contain the same imaging data, multiple detections at same locations were devised as a fail-safe strategy to provide enhanced robustness. This is analogous to a human observing an object from multiple perspectives to confirm a hypothesis. We observed suboptimal performance in the vertical scans, because those are reconstructed synthetically from horizontally acquired data. Any untracked motion between adjacent horizontal scans created artifacts in synthetically derived vertical B-scans, making the detection of EZ loss more challenging to the algorithm. The combined network outperformed the EZ loss map from individual networks, improving the overall robustness of the method. The 2-stage network efficiently learned an accurate model using the limited number of training examples available in our data set. At the first stage, a 2D network benefitted from a robust, powerful 2D

![Figure 7](image)

*Figure 7. Graph showing ellipsoid zone (EZ) loss area distributions for affected and unaffected groups detected with the automatic algorithm.*

![Figure 8](image)

*Figure 8. A, Scatterplot showing the relationship between ellipsoid zone (EZ) loss area and Humphrey visual field (HVF) mean deviation (MD). B, Scatterplot showing visual acuity as a function of closest distance to EZ loss from the fovea.*
network architecture and a larger number of training samples available from the B-scans. The determination of the EZ loss per volume at the second stage aggregating the output from already trained 2D slice detections was efficient compared with using a memory-intensive 3D network architecture.\textsuperscript{32} The algorithm successfully detected EZ loss in variable levels of SNR present in acquired images and did not exhibit any degradation of performance within the SNR range of 17.8 to 31.1 dB.

Our approach to detecting areas of EZ loss via instance segmentation (object detection followed by pixel segmentation) is algorithmically different from obtaining EZ loss after direct pixel labelling of retinal layers. Although both approaches are capable of yielding pixelwise segmentations of areas corresponding to EZ loss and producing en face EZ loss maps, we observed that detection-based segmentation yielded superior performance in the experiments. Instance detection before segmentation (instance-first strategy in the M-RCNN) can be advantageous in first selecting candidate regions with probable loss before labelling specific pixels depicting layer deterioration. In contrast, layer segmentation methods attempt to assign a label to all pixels in a B-scan where most do not exhibit EZ loss, and most of the pixels do not pertain to the EZ layer. When directly segmenting regions with EZ loss, this could lead to a challenging class imbalance problem where the number of negative pixels vastly outnumber the number of positive pixels in the data set. Thus, the instance-first strategy could be more efficient in learning with limited data. Although the results of applying a retinal layer segmentation demonstrated the potential to identify areas of EZ loss successfully, the measured performance was inferior to detection-based segmentation (Fig 4), at least without significant retraining using images from patients who use hydroxychloroquine. Furthermore, detection-based methods in principle can separate between different instances of the class (i.e., multiple instance of EZ loss), whereas the layer segmentation does not differentiate between multiple instances directly in its output. We did not use this additional feature because our objective was to quantify the entire region of EZ loss in a given eye.

The reported algorithm for automatic EZ loss detection could facilitate screening for retinal toxicity with the ability to quantify structural alterations in an objective, time-efficient, and cost-effective manner. The patients in this case-control study exhibited a wide range of severities of EZ deterioration, and the deep learning algorithm successfully detected and quantified EZ loss areas over a considerable range. The en face EZ loss map can provide both quantitative and qualitative insights such as the total EZ loss area, central fovea involvement, and topographical distribution of EZ loss. Toxicity determination based on EZ loss in SD OCT corroborated determinations of toxicity based on AAO-recommended screening guidelines, except in a few borderline cases. Although conspicuous EZ loss is an important criterion in determining toxicity according to our current understanding, multiple other metrics, including OCT metrics such as retinal layer thickness and intensity-based measurements, could augment multiple criteria to detect toxicity reliably. Ellipsoid zone attenuation without conspicuous loss also could provide evidence for early-stage changes. In our preliminary attempts, the reliable manual annotation of attenuated EZ loss was found to be highly variable and subjective, and thus, we did not include such examples (mostly in patients with < 100 \( \mu \text{m} \) of EZ loss) in our training data set. The additional metrics derived from SD OCT could be combined with the evidence from other functional tests (i.e., visual fields, mfERG) to make the final toxicity determination for the patient per AAO guidelines or could be used alone when necessary in cases where reliable functional testing is not feasible. Given that the screening recommendations call for annual screening, this may help to alleviate the need for clinical functional testing in some patients, especially where time, patient ability, or other obstacles to testing occur.

Subtleties and complexities exist in determination of toxicity during early stages when multiple screening methods convey discordant evidence, and debate remains as to which screening test—visual field testing, SD OCT, or mfERG—is the most accurate and sensitive in detecting the earliest evidence of retinal damage. In our data set, ancillary testing indicated toxicity in a small number of patients (9/168), although the OCT scan did not demonstrate any evidence of measurable EZ loss. Although ongoing research and studies (some involving longitudinal monitoring of patients) may reveal additional insights into the pathophysiological sequence of events, objective, quantifiable SD OCT-derived metrics can identify instances of early disease with detection of very small areas of EZ loss. Thus, the automatic method to EZ loss detection presented in this work provides an efficient, objective, quantitative method that can be incorporated into ongoing work that aims to refine criteria to detect toxicity both accurately and at the earliest onset.

Accuracy validation and machine learning model development are limited by the variability of humans in annotating ground truth. This variability in human annotations is an important limiting factor that challenges the learning of a generalized model with a limited number of training images. Our results indicate that the error was approaching that of human variability and that the algorithm could be improved further with a better understanding of early-stage structural alterations induced by retinal toxicity. Even at the level of performance reported herein, the algorithm has advantages in precision and repeatability to mitigate the subjective, qualitative assessment in current clinical practice.

Although the OCT scan protocol in this work covered a 8.9 \( \times \) 7.4-mm\(^2\) (768 \( \times \) 496 pixels) region, even wider OCT scanning protocols can have advantages, especially in Asian patients.\textsuperscript{33} In the future, the algorithm we have developed could be retrained and updated with additional examples to detect EZ loss successfully, accounting for different OCT scanning protocols. Additionally, our study is limited by a single-center study sample and images acquired by a single vendor machine. Because of the 10-fold cross validation strategy in which images of test patients were not biased by the training patients’ images, the reported performance reflects the accuracy attainable in a single-vendor setting with a limited data set. These results
warrant further investigation to assess the performance in multivendor and multicenter settings. The proposed framework could be translated easily to these settings and could benefit from ground truth from additional data sets to improve generalizability. Ongoing and future work include measuring the EZ loss changes longitudinally to monitor disease progression and correlating these structural measurements with the outcomes in functional tests to streamline further criteria for early and definitive diagnosis of toxicity.

Clinical translation of this tool would enable automated and objective identification of patients who demonstrate changes concerning for toxicity, which could aid the screening ophthalmologist. Corroborating these results with ancillary functional testing, identifying patients who would benefit from referral to specialists, or both would improve current screening methods. Implementation of this automatic algorithm also could help the feasibility of screening outside of ophthalmology offices as OCTs become more ubiquitous in internal medicine settings. Quantitative data produced from the algorithms could provide surrogate end points for use in clinical trials and intervention studies aimed at halting progression of degenerative changes. Furthermore, this tool could translate directly to other diseases of the outer retina that share structural features of loss of the EZ reflectivity band and could provide useful outcome measures in developing any therapeutics that may alter the course of the disease.

Footnotes and Disclosures

Originally received: April 20, 2021.
Final revision: September 15, 2021.
Accepted: September 16, 2021.
Available online: September 25, 2021. Manuscript no. D-21-00070.

1 National Eye Institute, National Institutes of Health, Bethesda, Maryland.
2 The National Center for Advancing Translational Sciences, National Institutes of Health, Bethesda, Maryland.

Presented at: Society of Photo-Optical Instrumentation Engineers (SPIE) Medical Imaging Conference, San Diego, California, February, 2021.

Disclosure(s):
All authors have completed and submitted the ICMJE disclosures form.
Emily Chew and Catherine Cukras, members of the editorial board of this journal, were recused from the peer-review process of this article and had no access to information regarding its peer-review.

The author(s) have no proprietary or commercial interest in any materials discussed in this article.

This research was supported by the Intramural Research Program of the NIH, National Eye Institute (Intramural Program grant no.: EY000498).

HUMAN SUBJECTS: Human subjects were included in this study. The human ethics committee at NIH approved the study. All research complied with the Health Insurance Portability and Accountability Act of 1996 and adhered to the tenets of the Declaration of Helsinki. All participants provided informed consent.

No animal subjects were included in this study.

Author Contributions:
Conception and design: De Silva, Hotaling, Chew, Cukras
Analysis and interpretation: De Silva, Hotaling, Chew, Cukras
Data collection: De Silva, Jayakar, Grisso, Cukras
Obtained funding: N/A; Study was performed as part of official duties as officers or employees of the US government. No additional funding was provided.

Overall responsibility: De Silva, Jayakar, Grisso, Hotaling, Chew, Cukras

Abbreviations and Acronyms:
AAO = American Academy of Ophthalmology; CPN = combined projection network; EZ = ellipsoid zone; IOU = intersection over union; mfERG = multifocal electroretinography; M-RCNN = mask region-based convolutional neural network; M-RCNNH = horizontal mask region-based convolutional neural network; M-RCNNV = vertical mask region-based convolutional neural network; SD = spectral-domain; SNR = signal-to-noise ratio; 3D = 3-dimensional; 2D = 2-dimensional.

Keywords:
Automatic detection, Deep learning, Ellipsoid zone loss, Hydroxychloroquine toxicity.

Correspondence:
Catherine A. Cukras, MD, PhD, Division of Epidemiology and Clinical Applications, National Eye Institute, National Institutes of Health, Bethesda, MD 20892. E-mail: cukrasc@nei.nih.gov.

References

1. Melles RB, Marmor MF. The risk of toxic retinopathy in patients on long-term hydroxychloroquine therapy. JAMA Ophthalmol. 2014;132:1453–1460.
2. Marmor MF, Hu J. Effect of disease stage on progression of hydroxychloroquine retinopathy. JAMA Ophthalmol. 2014;132:1105–1112.
3. Allahdina AM, Chen KG, Alvarez JA, et al. Longitudinal changes in eyes with hydroxychloroquine retinal toxicity. Retina. 2019;39:473–484.
4. Marmor MF, Kellner U, Lai TYY, et al. Recommendations on screening for chloroquine and hydroxychloroquine retinopathy (2016 revision). Ophthalmology. 2016;123:1386–1394.
5. Sadda SR, Chakravarthi U, Birch DG, et al. Clinical endpoints for the study of geographic atrophy secondary to age-related macular degeneration. Retina. 2016;36:1806–1822.
6. Birch DG, Bennett LD, Duncan JL, et al. Long-term follow-up of patients with retinitis pigmentosa receiving intraocular ciliary neurotrophic factor implants. Am J Ophthalmol. 2016:170:10–14.
7. Spadea RF, Cercio CA. Anatomical correlates to the bands seen in the outer retina by optical coherence tomography: literature review and model. Retina. 2011;31:1609–1619.
8. Pauleikhoff D, Bonelli R, Dubis AM, et al. Progression characteristics of ellipsoid zone loss in macular telangiectasia type 2. Acta Ophthalmol. 2019;97:e998–e1005.
9. Cai CX, Light JG, Handa JT. Quantifying the rate of ellipsoid zone loss in Stargardt disease. Am J Ophthalmol. 2018;186:1–9.
10. Sadda SR, Guymer R, Holz FG, et al. Consensus definition for atrophy associated with age-related macular degeneration on
OCT: Classification of Atrophy report 3. *Ophthalmology*. 2018;125:537–548.

11. Cai CX, Locke KG, Ramachandran R, et al. A comparison of progressive loss of the ellipsoid zone (EZ) band in autosomal dominant and X-linked retinitis pigmentosa. *Invest Ophthalmol Vis Sci*. 2014;55:7417–7422.

12. Ahn SJ, Joung J, Lee BR. En face optical coherence tomography imaging of the photoreceptor layers in hydroxychloroquine retinopathy. *Am J Ophthalmol*. 2019;199:71–81.

13. Ugwuegbu O, Uchida A, Singh RP, et al. Quantitative assessment of outer retinal layers and ellipsoid zone mapping in hydroxychloroquine retinopathy. *Br J Ophthalmol*. 2019;103:3–7.

14. Itoh Y, Vasani A, Ehlers JP. Volumetric ellipsoid zone mapping for enhanced visualisation of outer retinal integrity with optical coherence tomography. *Br J Ophthalmol*. 2016;100:295–299.

15. Marmor MF. Comparison of screening procedures in hydroxychloroquine toxicity. *Arch Ophthalmol*. 2012;130:461–469.

16. Loo J, Clemons TE, Chew EY, et al. Beyond performance metrics. *Ophthalmology*. 2020;127:799–801.

17. Petzold A, Balcer L, Calabresi PA, et al. Retinal layer segmentation in multiple sclerosis: a systematic review and meta-analysis. *Lancet Neurol*. 2017;16:797–812.

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. Etheridge T, Dobson ETA, Wiedenmann M, et al. A semi-automated machine-learning based workflow for ellipsoid zone analysis in eyes with macular edema: SCORE2 pilot study Vavvas DG. *PloS One*. 2020;15:e0232494.

20. Wang Y-Z, Galles D, Klein M, et al. Application of a deep machine learning model for automatic measurement of EZ width in SD-OCT images of RP. *Transl Vis Sci Technol*. 2020;9:15.

21. Hariri AH, Zhang HY, Ho A, et al. Quantification of ellipsoid zone changes in retinitis pigmentosa using en face spectral domain—optical coherence tomography. *JAMA Ophthalmol*. 2016;134:628.

22. Vyshnav MT, Sowmya V, Gopalakrishnan EA, et al. Deep learning based approach for multiple myeloma detection. In: 2020 11th International Conference on Computing, Communication and Networking Technologies (iCCCNt). IEEE. 2020:1–7. Available at: https://ieeexplore.ieee.org/document/9225651/. Accessed August 9, 2021.

23. Minatel PG, Oliveira BC, Albertazzi A. Comparison of Unet and Mask R-CNN for impact damage segmentation in lock-in thermography phase images. In: Beyerer J, Heizmann M, eds. *Automated Visual Inspection and Machine Vision IV*. SPIE; 2021:26. Available at: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11787/2600734/Comparison-of-Unet-and-Mask-R-CNN-for-impact-damage/10.1117/12.2600734.full. Accessed August 9, 2021.

24. Zhao T, Yang Y, Niu H, et al. Comparing U-Net convolutional networks with fully convolutional networks in the performances of pomegranate tree canopy segmentation. In: Larar AM, Suzuki M, Wang J, eds. *Multispectral, Hyperspectral, and Ultraspectral Remote Sensing Technology, Techniques and Applications VII*. SPIE; 2018:64. Available at: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10780/2325570/Comparing-U-Net-convolutional-networks-with-fully-convolutional-networks-in/10.1117/12.2325570.full. Accessed August 9, 2021.

25. Guo X, Wang F, Teodorov G, et al. Liver steatosis segmentation with deep learning methods. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). IEEE. 2019:24–27. Available at: https://ieeexplore.ieee.org/15354086. Accessed August 9, 2021.

26. Anderson C, Blaha GR, Marx JL. Humphrey visual field findings in hydroxychloroquine toxicity. *Eye*. 2011;25:1535–1545.

27. Pfau M, Von Der Emde L, De Sisternes L, et al. Progression of photoreceptor degeneration in geographic atrophy secondary to age-related macular degeneration. *JAMA Ophthalmol*. 2020;138:1026–1034.

28. Tsang AC, Ahmadi S, Hamilton J, et al. The diagnostic utility of multifocal electroretinography in detecting chloroquine and hydroxychloroquine retinal toxicity. *Am J Ophthalmol*. 2019;206:132–139.

29. He K, Gkioxari G, Dollár P, Girshick R, Mask R-CNN. *IEEE Trans Pattern Anal Mach Intell*. 2020;42:386–397.

30. Ren S, He K, Girshick R, Sun J, Faster R-CNN. Towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Mach Intell*. 2017;39:1137–1149.

31. Chen Y, Li J, Xiao H, et al. Dual path networks. In: Advances in Neural Information Processing Systems. 2017:4467–4475.

32. Siewersden JH, Levine M, De Silva TS, et al. Automatic vertebrae localization in spine CT: a deep-learning approach for image guidance and surgical data science. In: Fei B, Linte CA, eds. *Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions, and Modeling*. SPIE; 2019:27. Available at: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10951/2513915/Automatic-vertebrae-localization-in-spine-CT-a-deep-learning/10.1117/12.2513915.full. Accessed February 10, 2021.

33. Ahn SJ, Joung J, Lim HW, Lee BR. Optical coherence tomography protocols for screening of hydroxychloroquine retinopathy in Asian patients. *Am J Ophthalmol*. 2017;184:11–18.

34. Csaky K, Ferris F, Chew EY, et al. Report from the NEI/FDA Endpoints Workshop on Age-Related Macular Degeneration and Inherited Retinal Diseases. *Invest Ophthalmol Vis Sci*. 2017;58:3456.