tvst

Retina

Prediction of Long-Term Treatment Outcomes for Diabetic Macular Edema Using a Generative Adversarial Network

Jiwon Baek[1–4], Ye He[1,4], Mehdi Emamverdi[1,4], Alireza Mahmoudi[1,4], Muneeswar Gupta Nittala[1], Giulia Corradetti[1,4], Michael Ip[1,4], and SriniVas R. Sadda[1,4]

[1] Doheny Eye Institute, Pasadena, CA, USA
[2] Department of Ophthalmology, Bucheon St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Bucheon, Gyeonggi-do, Republic of Korea
[3] Department of Ophthalmology, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea
[4] Department of Ophthalmology, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA

Correspondence: SriniVas R. Sadda, Doheny Eye Institute, Department of Ophthalmology, David Geffen School of Medicine at UCLA, 1355 San Pablo Street, Los Angeles, CA 90033, USA. e-mail: ssadda@doheny.org

Received: March 6, 2024
Accepted: May 25, 2024
Published: July 3, 2024

Keywords: diabetic macular edema (DME); generative adversarial network (GAN); prediction; randomized controlled trial (RCT); anti-vascular endothelial growth factor (VEGF)

Citation: Baek J, He Y, Emamverdi M, Mahmoudi A, Nittala MG, Corradetti G, Ip M, Sadda SR. Prediction of long-term treatment outcomes for diabetic macular edema using a generative adversarial network. Transl Vis Sci Technol. 2024;13(7):4, https://doi.org/10.1167/tvst.13.7.4

Purpose: The purpose of this study was to analyze optical coherence tomography (OCT) images of generative adversarial networks (GANs) for the prediction of diabetic macular edema after long-term treatment.

Methods: Diabetic macular edema (DME) eyes ($n = 327$) underwent anti-vascular endothelial growth factor (VEGF) treatments every 4 weeks for 52 weeks from a randomized controlled trial (CRTH258B2305, KINGFISHER) were included. OCT B-scan images through the foveal center at weeks 0, 4, 12, and 52, fundus photography, and retinal thickness (RT) maps were collected. GAN models were trained to generate probable OCT images after treatment. Input for each model were comprised of either the baseline B-scan alone or combined with additional OCT, thickness map, or fundus images. Generated OCT B-scan images were compared with real week 52 images.

Results: For 30 test images, 28, 29, 15, and 30 gradable OCT images were generated by CycleGAN, UNIT, Pix2PixHD, and RegGAN, respectively. In comparison with the real week 52, these GAN models showed positive predictive value (PPV), sensitivity, specificity, and kappa for residual fluid ranging from 0.500 to 0.889, 0.455 to 1.000, 0.357 to 0.857, and 0.537 to 0.929, respectively. For hard exudate (HE), they were ranging from 0.500 to 1.000, 0.545 to 0.900, 0.600 to 1.000, and 0.642 to 0.894, respectively. Models trained with week 4 and 12 B-scans as additional inputs to the baseline B-scan showed improved performance.

Conclusions: GAN models could predict residual fluid and HE after long-term anti-VEGF treatment of DME.

Translational Relevance: The implementation of this tool may help identify potential nonresponders after long-term treatment, thereby facilitating management planning for these eyes.

Introduction

Diabetic macular edema (DME) stands as a prominent cause of global blindness and currently represents the second-largest segment within the retinal disease treatment landscape following age-related macular degeneration (AMD).[1]

For its treatment, anti-vascular endothelial growth factor (VEGF) injection is the primary option. A firm basis for the treatment of DME with anti-VEGF has been established by numerous randomized controlled trials (RCTs) conducted during the last 2 decades, including RISE and RIDE, VIVID and VISTA, and DRCR.net protocols I and T.[2–5]

Nonetheless, approximately 20% to 40% eyes with DME are reported to be refractory to monthly intravitreal VEGF monotherapy in these landmark clinical trials. The evaluation of imaging biomarkers has been one of the important approaches used to predict the outcome after anti-VEGF treatment for DME.[2] Studies have demonstrated that intraretinal cystoid fluid, total retinal thickness (RT), outer nuclear layer, subretinal fluid (SRF), and hyper-reflective foci (HF) on optical coherence tomography (OCT) can serve as markers indicating a higher probability of suboptimal treatment response to anti-VEGF therapy in DME eyes.[6,7]

Predicting the response to anti-VEGF treatment in DME has previously been attempted with machine learning (ML) methods.[8–12] Cao et al. presented compelling results with a random forest model, utilizing features extracted from OCT, to predict the response to anti-VEGF treatment in DME eyes after a 3-month period.[8] Other studies by Chen et al. and Gallardo et al. demonstrated the efficacy of ML models trained with baseline clinical and OCT characteristics for forecasting the longer-term prognosis in DME eyes.[9,13] These prior investigations underscore the viability of utilizing baseline or initial clinical and OCT features in predicting the prognosis of DME in ML models. The application of generative adversarial network (GAN) for predicting the prognostic morphology of OCT images is a recent introduction to this line of research.

Given these observations, GANs can visualize prognosis of the disease by training to generate possible post-treatment image from baseline OCT or other images with additional biomarkers.[14–17] Long-term follow-up is essential for patients with DMEs, and the value of predicting treatment outcomes, thereby identifying potential nonresponders, remains significant for individuals affected by DME. Unlike traditional ML models that only show binary prediction results, GAN models can show the predicted results in images, allowing for a more intuitive understanding of the disease progression. Here, we trained and evaluated the performance of GAN models in prediction of long-term images for DME based on data after regular injections for 1 year.

Methods

Study Subjects

This study utilized data collected during the KINGFISHER study (CRTH258B2305), a randomized, double-masked, multicenter phase III study assessing the efficacy and safety of brolucizumab (Beovu, Novartis, Basel, Switzerland) every 4 weeks versus aflibercept every 4 weeks (Eylea, Bayer, Leverkusen, Germany) in patients with DME. Detailed descriptions of the KINGFISHER study can be found at: https://clinicaltrials.gov/study/NCT03917472. Informed consent was obtained from all subjects before enrollment. Institutional review board approval was obtained by Doheny Eye Institute (IRB protocol number 15-000086) for these post hoc analyses. The research followed the tenets of the Declaration of Helsinki for research involving human subjects.

The inclusion criteria for the study are as follows: (1) patients ≥ 18 years of age at baseline; (2) patients with type 1 or type 2 diabetes mellitus (DM) and hemoglobin A1c ≤ 12% at screening; (3) study eye visual impairment due to DME, with a best corrected visual acuity (BCVA) score between 73 and 23 letters using Early Treatment Diabetic Retinopathy Study (ETDRS) visual acuity testing charts at both screening and baseline; and (4) DME involving the center of the macula, with central subfield thickness ≥ 320 µm on OCT.

The key exclusion criteria for the study are as follows: (1) high-risk proliferative diabetic retinopathy (PDR) in the study eye as per investigator assessment at both screening and baseline; (2) concomitant conditions or ocular disorders in the study eye at screening or baseline, which may confound interpretation of the study results (e.g. structural damage of the fovea, vitreous hemorrhage, retinal detachment, retinal vein/arterial occlusion, neovascularization of iris or choroidal neovascularization of any cause, uncontrolled glaucoma, and amblyopia); (3) any active intraocular or periocular infection or active intraocular inflammation in either eye at screening or baseline; (4) use of anti-VEGF therapies, intraocular surgery, or laser photocoagulation in the study eye during the 3-month period prior to baseline; and (5) use of intraocular corticosteroids, including dexamethasone and fluticasone implant in the study eye during the 6-month period prior to baseline.

Image Collection and GAN Model Training

Of the 517 total patients enrolled, only those who completed the 52-week follow-up and underwent OCT volume scan (20 × 20 degrees; 512 A-scans × 97 B-scans) imaging with Spectralis (Spectralis; Heidelberg Engineering, Heidelberg, Germany) were eligible for this analysis.

From the selected patients, OCT volume raw data at baseline and weeks 4, 12, and 52 were saved as E2E files and OCT B-scans at the fovea were extracted and saved as JPG files (1024 × 496 pixels). In addition, the infra-red fundus photographs (FPs; 434 × 343 pixels) and RT OCT heatmap (794 × 794 pixels) were collected from each baseline E2E file and saved in JPG files. The RT heatmap was generated from automated segmenta-

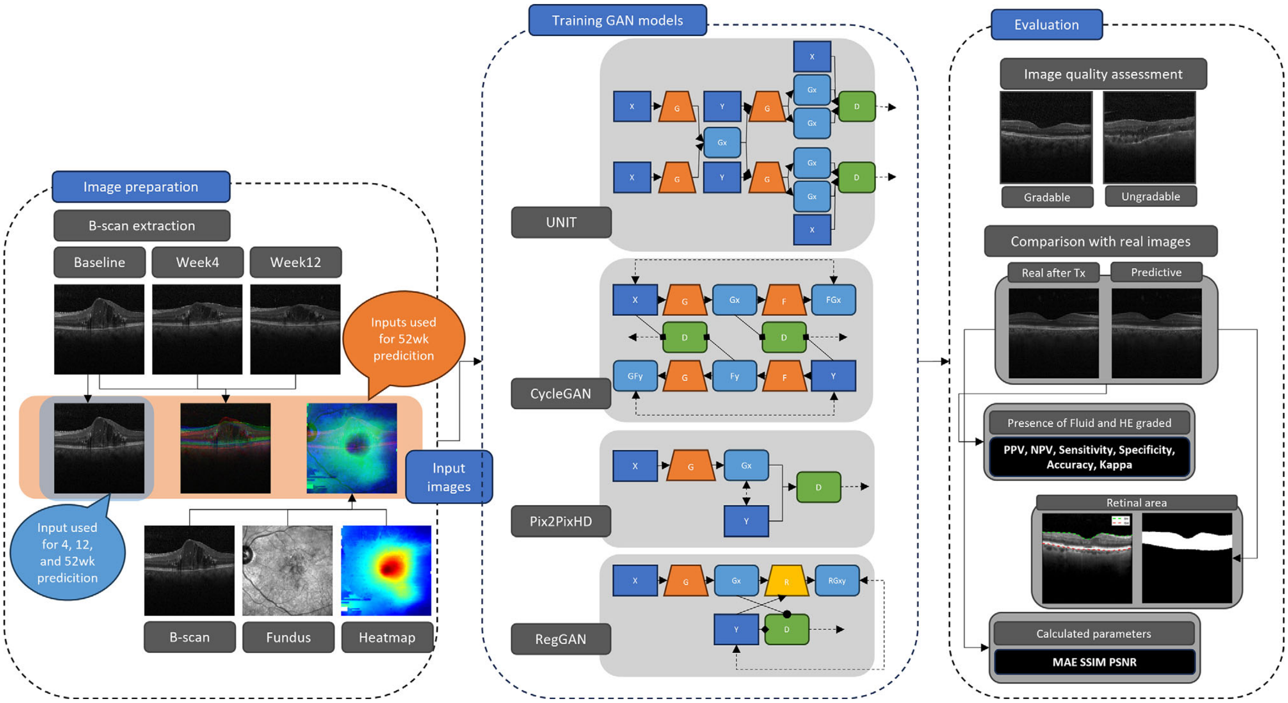

Figure 1. Image preparation, generative adversarial network (GAN) model training, and evaluation. (Image preparation) Optical coherence tomography (OCT) B-scans at the fovea at the baseline, weeks 4, 12, and 52, and fundus photography (FP), and retinal thickness (RT) heatmap at the baseline were collected. Baseline B-scan was used as input for generating images for weeks 4, 12, and 52. In addition, B-scans of baseline + week 4 + week 12, and B-scan + FP + RT heatmap, were used as inputs for generating week 52 image. The dataset was randomly divided into a training set ($n = 297$) and a test set ($n = 30$). (Training GAN models) Pix2PixHD, UNIT, CycleGAN, and RegGAN were used for training on generating probable post-treatment OCT B-scan images in the training set. (Evaluation) Generated images from the test set were first categorized as gradable or ungradable. Then, the presence of fluid and hard exudate (HE) in each individual image was graded by experts, and the grading results were compared with the ground truth. Retinal area defined as an area between the internal limiting membrane and retinal pigment epithelial line was also compared between generated image and the ground truth. Mean absolute error (MAE), peak signal-to-noise ratio (PSNR), and structural similarity index (SSIM) were calculated for each GAN model.

tion information stored in the raw file by the Spectralis system. All images were resized to 256 × 256 pixels.

Of a total of 327 cases, the dataset was randomly divided into a training set ($n = 297$) and a test set ($n = 30$; Fig. 1). Pix2PixHD, UNIT, CycleGAN, and RegGAN were used for training and generation of post-treatment OCT B-scan images. Using baseline B-scans as input images, the GAN models were trained to generate probable post-treatment OCT images at weeks 4, 12, and 52. For the generation of OCT images at week 52, concatenated images of B-scans of baseline, week 4, and week 12, and concatenated images of the baseline B-scan, FP, and RT heatmap were used as inputs in addition to B-scans only. Real 52 week OCT B-scan images were used as reference for transferred image for training. To train RegGAN, CycleGAN, and UNIT, the default hyperparameters from the original networks were used (i.e. learning rate adjusted to 0.0001, with the number of iterations to linearly decay the learning rate to zero adjusted to 20, and the total number of epochs adjusted to 200).[18] For Pix2PixHD, the initial learning rate was 0.0002, the number of iteration to linearly decay learning rate to zero was set to 100, and epoch was set to 100 (total 200 epochs).

All experiments were performed with the PyTorch deep learning framework (version 2.1.0 + cu121) in Python (version 3.11.5; Python Software Foundation, Wilmington, DE, USA) using NVIDIA RTX3090 (NVIDIA, Santa Clara, CA, USA), and Intel i7 CPU 3.6-GHz processor.

GAN Model Test and Evaluation

Presumable post-treatment images based on input images were generated by the trained GAN models using the baseline images from the test set. OCT images were assessed by two retinal specialists (authors YH and ME) with over 5 years of ophthalmology experience. First, a mixture of real and generated/synthetic OCT images was given to the blinded graders to categorize each image as gradable or ungradable. Images were considered gradable if they exhibited identifiable and undistorted retinal structures and layers, allowing graders to accurately identify the types of retinal lesions present. Then, the graders were asked to identify the presence of any fluid (i.e. subretinal and/or intra-retinal) and hard exudate (HE) in each individual image and record it as present or none. In case of disagreement between the graders, a third grader (author JB) was consulted to provide a grading, and a final decision was made reflecting the majority opinion. The grading results were compared between the generated image and the ground truth, which was the real OCT B-scan image at week 52. The retinal area for each OCT B-scan was obtained in a semi-automated manner. The internal limiting membrane and retinal pigment epithelial line were automatically detected on each B-scan using a graph-based segmentation of retinal areas.[19,20] Erroneous segmentations were corrected by changing the contrast and brightness of the image, and five images (8%) required manual correction. The retinal area obtained as pixels from each model was compared to that of the ground truth. Additionally, the mean absolute error (MAE), peak signal-to-noise ratio (PSNR), and structural similarity index (SSIM) were calculated for each GAN model (see Fig. 1).[21]

Statistical Analysis

Statistical analysis was performed using SPSS version 26.0.1 for Windows (IBM Corp., Armonk, NY, USA) and MATLAB (R2023a; The MathWorks, Inc., Natick, MA, USA). Intergrader reliability regarding the presence of lesions was assessed by calculating Cohen's kappa. To evaluate the performance of GAN models, we calculated the positive predictive value (PPV), negative predictive value (NPV), sensitivity, specificity, accuracy, and Cohen's kappa for the graded presence of fluid and HE between the ground truth and generated images. The Kruskal-Wallis test was used to compare the mean performance variables among models. The Wilcoxon signed-rank test was used to compare the retinal area between each model and the ground truth. A P value < 0.05 was considered significant.

Results

Qualitative Assessment for Generated Post-Treatment Images for Week 52

Week 52 post-treatment images were generated by Pix2PixHD, CycleGAN, UNIT, and RegGAN from baseline OCT B-scans. With regard to the frequency of gradable images, 15 of 30 (50%) were deemed gradable for Pix2PixHD, 28 of 30 (93%) for CycleGAN, 29 of 30 (97%) for UNIT, and 30 of 30 (100%) for RegGAN ($P < 0.001$). Confabulation of the generated images can be observed in Pix2PixHD models with high frequency. The kappa value for OCT grading was 0.835 between the 2 graders. Examples of the week 52 post-treatment images generated by each GAN model are shown in Figure 2.

Prediction of Fluid and HE by GAN Models Using Baseline B-Scan at Weeks 4, 12, and 52

In the test set, the number of eyes with residual fluid at weeks 4, 12, and 52 were 16, 16, and 11, respectively ($P = 0.957$). HE was observed in 12, 10, and 11 eyes at weeks 4, 12, and 52. The kappa value for ground-truth fluid and HE as determined by the graders was 0.814 and 0.817, respectively.

CycleGAN, UNIT, and RegGAN were trained to generate week 4, 12, and 52 post-treatment OCT images from baseline OCT B-scans. The ranges of PPV, NPV, sensitivity, specificity, accuracy, and kappa for residual fluid were 0.500 to 0.889, 0.556 to 1.000, 0.455 to 1.000, 0.357 to 0.857, 0.567 to 0.933, and 0.537 to 0.929, respectively, and those for HE were 0.500 to 1.000, 0.773 to 0.944, 0.545 to 0.900, 0.600 to 1.000, 0.667 to 0.900, and 0.642 to 0.894, respectively (Table 1).

No significant difference was observed among weeks 4, 12, and 52 in terms of PPV, NPV, sensitivity, specificity, accuracy, and kappa for residual fluid and HE (all $P > 0.05$; Fig. 3). Among the GAN models, RegGAN showed the highest NPV, specificity, accuracy, and kappa mean (all $P \leq 0.05$; Fig. 4).

Prediction of Fluid and HE by GAN Models Using Multiple Input Images at Week 52

When week 4 and week 12 OCT B-scans were added to the baseline OCT B-scan as inputs, the ranges of PPV, NPV, sensitivity, specificity, accuracy, and kappa for residual fluid were 0.909 to 1.000, 0.826 to 0.950, 0.636 to 0.909, 0.947 to 1.000, 0.867 to 0.967, and 0.858 to 0.964, respectively, and those for HE were 0.750 to 1.000, 0.773 to 0.947, 0.545 to 0.909, 0.895 to 1.000, 0.767 to 0.933, and 0.752 to 0.929, respectively (Table 2).

When the FP and RT heatmap were added to the baseline OCT B-scan as inputs, the ranges of PPV, NPV, sensitivity, specificity, accuracy, and kappa for residual fluid were 0.563 to 0.800, 0.850 to 0.929, 0.727 to 0.909, 0.632 to 0.895, 0.700 to 0.833, and 0.678 to 0.822, respectively; and those for HE were 0.750 to 1.000, 0.773 to 0.900, 0.545 to 0.818, 0.895 to 1.000, 0.767 to 0.900, and 0.752 to 0.893, respectively (see

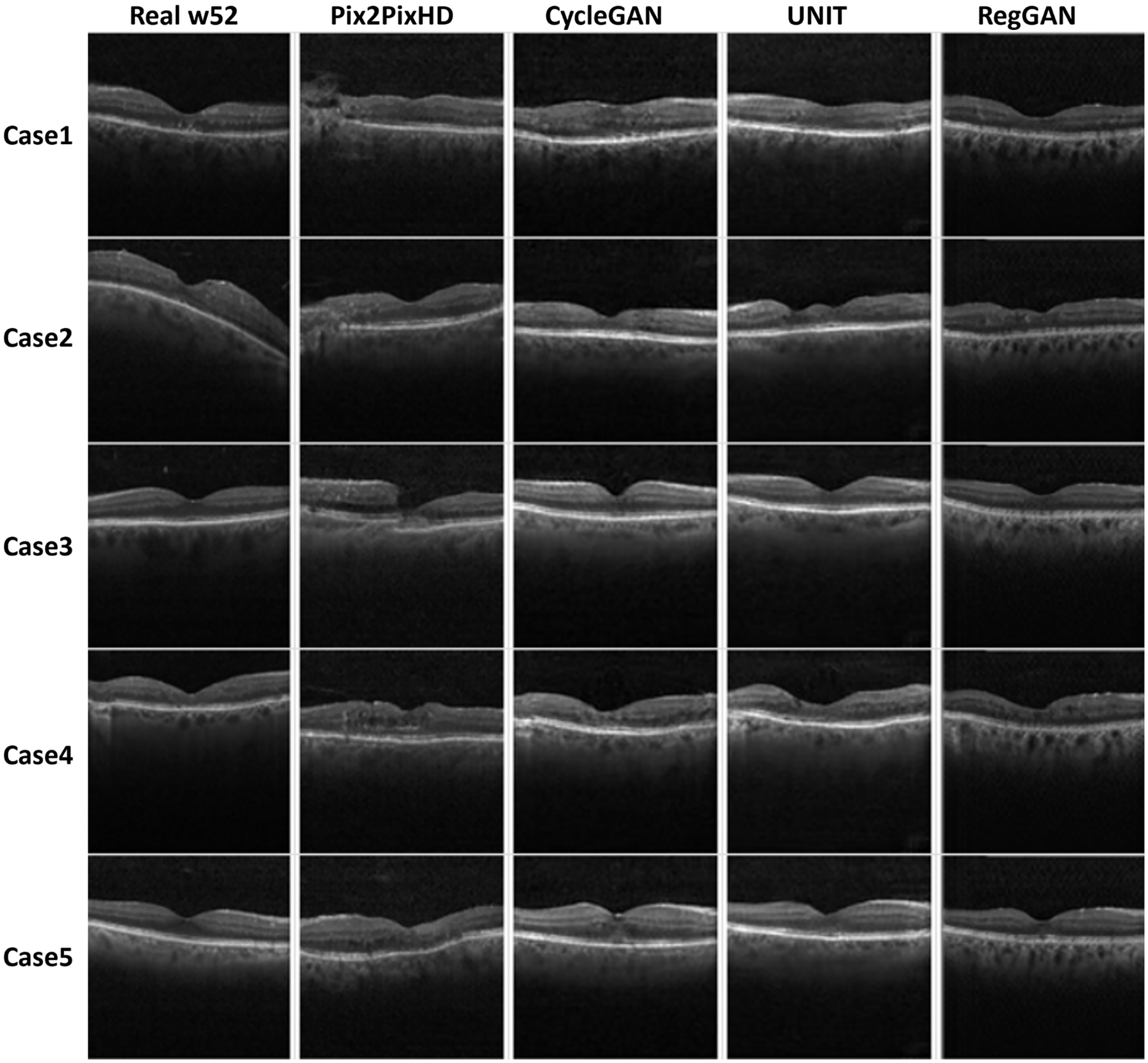

Figure 2. Examples of week 52 post-treatment images generated by each generative adversarial network (GAN) model. The first column represents the real week 52 post-treatment images, which serve as the ground truth. The corresponding examples generated using each GAN model and test set data are presented in the same row from the second to the fifth column (Pix2PixHD, CycleGAN, UNIT, and RegGAN from left to right). Confabulation of the generated images can be observed in Pix2PixHD models with high frequency, as seen in the first three rows.

Table 2). PPV, specificity, accuracy, and kappa mean were higher in additional week 4 and week 12 OCT B-scans input model compared to other input models (Fig. 5).

Comparison of Retinal Area Between Generated Images and Real Week 52 Images

For the ground truth OCT B-scan 256 × 256 pixels image, the mean retinal area in pixels was 2049.24 ± 1.00. The mean retinal pixel area for the CycleGAN baseline B-scan only, the baseline B-scan with week 4 added, the baseline B-scan with week 12 added, and the fundus and RT included models were 11437.6 ± 1430.53, 11492.63 ± 2030.99, and 12176.4 ± 1827.23, respectively. For the RegGAN models, those were 11562.27 ± 1081.83, 11299.07 ± 760.27, and 11687.60 ± 1039.92, and were 11254.3 ± 1028.68, 12062.3 ± 2602.36, 11997.73 ± 1725.57, and 11556 ± 2049.24 for the UNIT models, respectively. There was no significant difference in the predicted retinal area for all models compared to the ground truth (all $P > 0.05$).

Table 1. PPV, NPV, Sensitivity, Specificity, Accuracy, and Kappa Values for Post-Treatment Prediction of IRF/SRF and HE by GAN Models Trained Using OCT B-Scans Only, in Comparison to Real OCT Images at Weeks 4, 12, and 52

	Week 4			Week 12			Week 52		
Post-Treatment	CycleGAN	UNIT	RegGAN	CycleGAN	UNIT	RegGAN	CycleGAN	UNIT	RegGAN
IRF/SRF									
PPV	0.643	0.571	0.889	0.700	0.722	0.833	0.533	0.500	0.692
NPV	0.563	0.556	1.000	0.800	0.750	0.917	0.800	0.700	0.882
Sensitivity	0.563	0.750	1.000	0.875	0.813	0.938	0.727	0.455	0.818
Specificity	0.643	0.357	0.857	0.571	0.643	0.786	0.632	0.737	0.789
Accuracy	0.600	0.567	0.933	0.733	0.733	0.867	0.667	0.633	0.800
Kappa mean	0.571	0.537	0.929	0.715	0.715	0.857	0.643	0.609	0.786
HE									
PPV	0.909	0.727	0.714	0.500	0.750	0.889	0.750	1.000	0.818
NPV	0.895	0.789	0.875	0.857	0.944	0.905	0.773	0.792	0.895
Sensitivity	0.833	0.667	0.833	0.800	0.900	0.800	0.545	0.545	0.818
Specificity	0.944	0.833	0.778	0.600	0.850	0.950	0.895	1.000	0.895
Accuracy	0.900	0.767	0.800	0.667	0.867	0.900	0.767	0.833	0.867
Kappa mean	0.893	0.751	0.786	0.642	0.858	0.894	0.752	0.823	0.858

GAN, generative adversarial network; IRF, intraretinal fluid; NPV, negative predictive value; OCT, optical coherence tomography; PPV, positive predictive value; SRF, subretinal fluid.

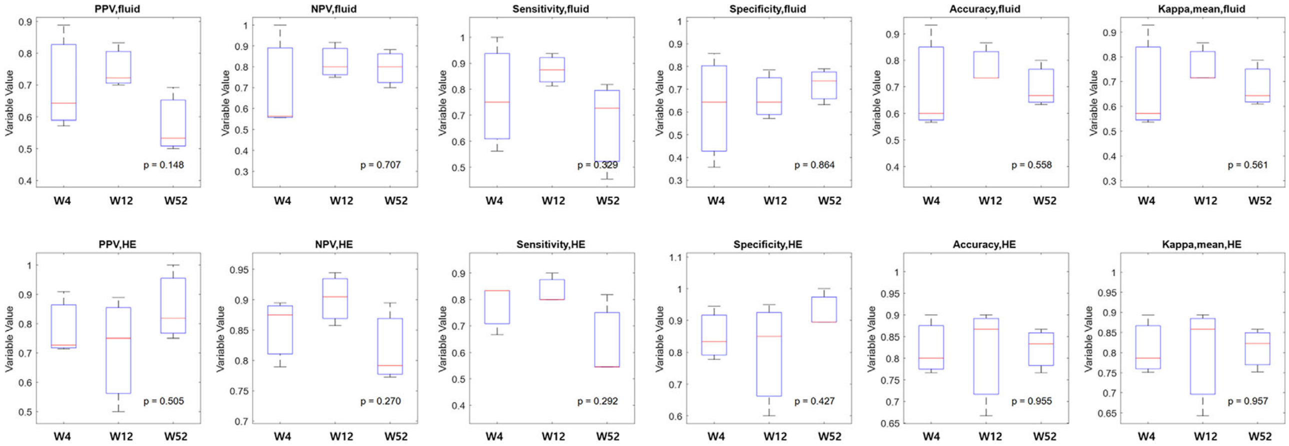

Figure 3. Comparison of mean performance parameters of the three GAN models (CycleGAN, UNIT, and RegGAN) for post-treatment prediction of residual fluids and HE at weeks 4, 12, and 52. (*Top row*) Prediction parameters for residual fluids. (*Bottom row*) Prediction parameters for HE. There was no significant difference in terms of PPV, NPV, sensitivity, specificity, accuracy, and kappa for presence of residual fluid and HE among prediction for weeks 4, 12, and 52 (all $P > 0.05$). GAN, generative adversarial network; HE, hard exudate; NPV, negative predictive value; PPV, positive predictive value.

MAE, PSNR, and SSIM for GAN Models

The ranges for MAE, PSNR, and SSIM by GAN models for 52 weeks compared to the ground truth were 16.840 to 26.803, 15.286 to 19.307, and 0.377 to 0.500, respectively. Compared to the baseline, those ranges were 3.229 to 26.478, 15.390 to 33.382, and 0.377 to 0.948, respectively (Table 3).

Discussion

In this study, we trained and evaluated the performance of GAN models for generating predictive OCT B-scan images after long-term anti-VEGF treatment in DME eyes. The results of the study showed that GAN models can not only generate acceptable quality OCT

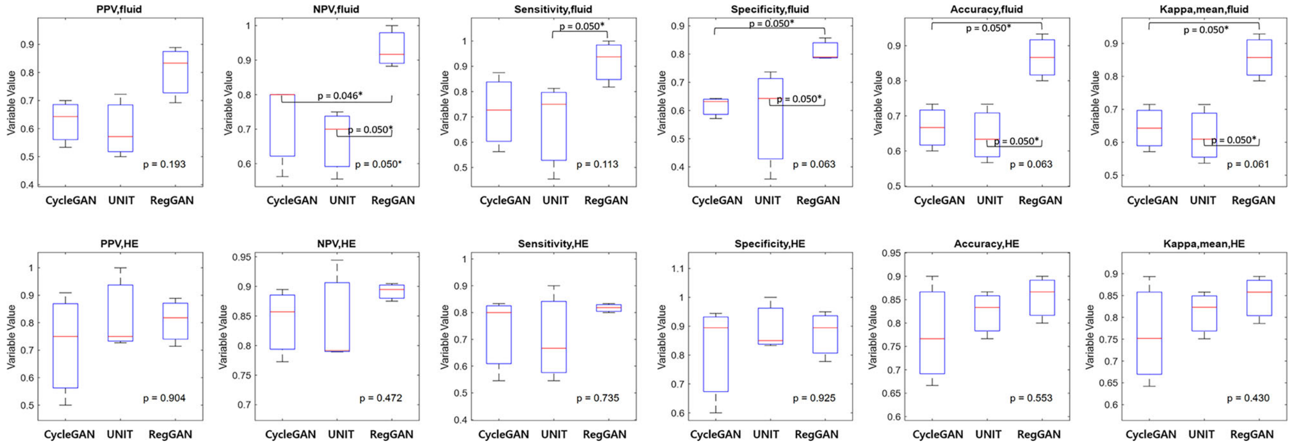

Figure 4. Comparison of mean performance parameters of weeks 4, 12, and 52 for post-treatment prediction of residual fluids and HE among models CycleGAN, UNIT, and RegGAN. (*Top row*) Prediction parameters for residual fluids. (*Bottom row*) Prediction parameters for HE. Among GAN models, RegGAN showed the highest NPV, specificity, accuracy, and kappa mean (all $P \leq 0.05$). GAN, generative adversarial network; HE, hard exudate; NPV, negative predictive value; PPV, positive predictive value. * Statistically significant *P* value.

Table 2. PPV, Sensitivity, and Specificity Values for Post-Treatment Prediction of IRF/SRF and HE by GAN Models Trained Using Multiple Input Images, in Comparison to Real OCT Images at Week 52

Post-Treatment	Features	Parameters	CycleGAN	UNIT	RegGAN
OCT B-scans (0, 4, 12 wk)	IRF/SRF	PPV	0.909	1.000	1.000
		NPV	0.947	0.826	0.950
		Sensitivity	0.909	0.636	0.909
		Specificity	0.947	1.000	1.000
		Accuracy	0.933	0.867	0.967
		Kappa mean	0.929	0.858	0.964
	HE	PPV	0.750	1.000	0.909
		NPV	0.773	0.792	0.947
		Sensitivity	0.545	0.545	0.909
		Specificity	0.895	1.000	0.947
		Accuracy	0.767	0.833	0.933
		Kappa mean	0.752	0.823	0.929
OCT B-scan + fundus + thickness map	IRF/SRF	PPV	0.625	0.563	0.800
		NPV	0.929	0.857	0.850
		Sensitivity	0.909	0.818	0.727
		Specificity	0.684	0.632	0.895
		Accuracy	0.767	0.700	0.833
		Kappa mean	0.750	0.678	0.822
	HE	PPV	0.750	1.000	0.900
		NPV	0.773	0.792	0.900
		Sensitivity	0.545	0.545	0.818
		Specificity	0.895	1.000	0.947
		Accuracy	0.767	0.833	0.900
		Kappa mean	0.752	0.823	0.893

GAN, generative adversarial network; IRF, intraretinal fluid; NPV, negative predictive value; OCT, optical coherence tomography; PPV, positive predictive value; SRF, subretinal fluid.

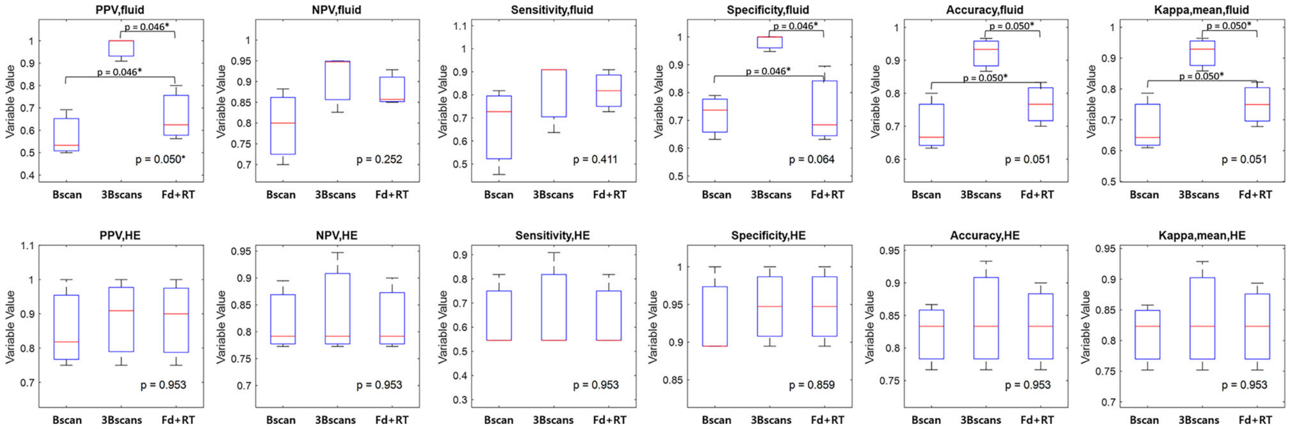

Figure 5. **Comparison of mean performance parameters of the three GAN models (CycleGAN, UNIT, and RegGAN) for post-treatment prediction of residual fluids and HE at week 52 among different inputs.** (*Top row*) Prediction parameters for residual fluids. (*Bottom row*) Prediction parameters for HE. The PPV, specificity, accuracy, and kappa mean were higher in weeks 4 and 12 OCT B-scans with added input model compared to other week 52 models ($P = 0.046$, 0.046, 0.05, and 0.05, and $P = 0.046$, 0.046, 0.05, and 0.05, respectively). GAN, generative adversarial network; HE, hard exudate; NPV, negative predictive value; PPV, positive predictive value; B-scan, baseline B-scan only; 3 B-scans, baseline + week 4 + week 12 B-scans; Fd + RT, baseline B-scan + fundus photography (FP) + retinal thickness (RT) map. * Statistically significant *P* value.

Table 3. MAE, PSNR, and SSIM for Predicted OCT Images at Weeks 4, 12, and 52 From GAN Models

		Compared to Week 52			Compared to Baseline		
Inputs	Parameters	CycleGAN	UNIT	RegGAN	CycleGAN	UNIT	RegGAN
Week 4	MAE	16.893	24.271	16.840	3.493	24.061	17.895
	PSNR	19.180	16.191	19.307	33.270	16.277	19.034
	SSIM	0.498	0.462	0.474	0.923	0.448	0.450
Week 12	MAE	24.541	25.825	18.503	3.229	25.545	19.298
	PSNR	16.318	16.091	18.527	33.382	16.151	18.403
	SSIM	0.450	0.432	0.446	0.948	0.423	0.426
Week 52	MAE	21.770	21.178	19.335	5.187	16.692	19.873
	PSNR	16.812	17.146	18.397	27.565	17.912	18.342
	SSIM	0.497	0.500	0.389	0.897	0.522	0.377
OCT B-scans (0, 4, 12 wk)	MAE	17.293	23.826	17.550	17.331	23.776	18.621
	PSNR	18.991	16.401	19.069	19.095	16.424	18.738
	SSIM	0.460	0.488	0.450	0.504	0.470	0.430
OCT B-scan	MAE	22.707	26.803	18.632	22.212	26.478	19.311
+ fundus	PSNR	16.840	15.286	18.545	17.020	15.390	18.422
+ thickness map	SSIM	0.496	0.492	0.428	0.481	0.479	0.410

GAN, generative adversarial network; MAE, normalized mean absolute error; OCT, optical coherence tomography; PSNR, peak signal to noise ratio; SSIM, structural similarity index.

images, but also can predict the presence of persistent fluid or HE after treatment in the long term.

GAN model selection was based on previous studies on GAN for other medical and retinal images.[14–16,22,23] The results for OCT generation and fluid prediction in the current study were comparable to previous studies. Lee et al. used conditional GAN to generate post-treatment OCT B-scan images for AMD.[14] They reported that the accuracy, specificity, and NPV for IRF, SRF, and pigment epithelial detachment were 77.0 to 91.9%, 94.1 to 95.1%, and 54.7 to 96.5%, respectively, which was comparable to our results on DME. Lui et al. compared the performances of several GAN models for the prediction of OCT appearance 1-month after an uncontrolled anti-VEGF treatment.[15] They reported that 92% of the images were difficult to differentiate from the real OCT images by retinal specialists, which was similar in the present study.

Of the GAN models used in this study, RegGAN showed the best performance in predicting residual fluid. This result is in accordance with a previous study by Lui et al.[15] We hypothesize that this is due to the more refined structure of the RegGAN model compared to the others. RegGAN has been updated from CycleGAN so that the generator is trained with an additional registration network to fit the misaligned noise distribution adaptively. Because the OCT images in the current study were not aligned between the baseline and week 52, this may explain why the RegGAN model yielded the best performance.

There were significant differences in the performance of GAN models among weeks 4, 12, and 52 for the prediction of residual fluid and HE. Just as other ML models designed for longer-term predictions have demonstrated effectiveness, the GAN models in the present study also proved to be as effective as for short-term predictions.

Features seen on fundus photography, such as retinal microaneurysms, hemorrhages, intraretinal microaneurysms, and venous beading, have been found to be associated with visual prognosis in diabetic retinopathy.[24] Adding FP and RT maps as input data, however, did not significantly enhance the performance of the GAN models. Although FP and RT maps may include important features that might be of relevance for predicting DME treatment outcomes, they are markedly different in structure compared to B-scan images. Using images with highly dissimilar structures as input images might not have helped to optimize the output of GAN. On the other hand, additional B-scans from weeks 4 and 12 improved the performance of GAN models in predicting B-scans results at week 52. This finding can be supported by several previous studies that have shown that consecutive initial BCVA or OCT features are good predictors for outcomes in patients with DME.[9,10]

An MAE of 26.74 ± 21.28 µm compared to the real image (327.67 ± 209.11 µm, 8.16% difference) was observed for central macular thickness (CMT) in the study by Liu et al.[15] It is unclear how they measured CMT in that study. As the present study utilized one central B-scan from the volume scan, the retinal area was calculated instead of the central subfield macular thickness. The result revealed no significant difference between the ground truth and each model, and the mean difference was only 2.35%. This indicates that the models used in the present study also performed well in estimating post-treatment retinal thickness.

Alignment differences between the baseline and real week 52 images used in training and testing might have contributed to SSIM values in the current study, which were below 0.6. SSIM is a method for measuring the similarity between two images. It compares the luminance, contrast, and structure of the images to determine their similarity.[21] A lower SSIM value indicates that the two images are less similar in terms of these structures. Whereas some OCT devices use a “Z-lock” during scan acquisition, the Spectralis does not, thereby causing some degree of differences in the Z-axis position within and between scans.[25] When SSIMs were calculated for images generated by CycleGAN and the baseline image, the value went up to 0.948. Unlike UNIT, which uses both baseline and week 52 images for synthetic image generation at the first generator, CycleGAN uses the baseline image as an input for the first generator and week 52 images are used for calculating adversarial loss by the discriminator, and the resemblance of the alignment is greater for the generated and baseline image in CycleGAN.

Our study is not without limitations. First, the size of the study sample was relatively small for GAN model training, which may have limited the maximum-achievable performance for these models. Additional information, including OCT angiography, fundus autofluorescence, cytokine profiles, and systemic chemical profiles, could also potentially yield better performance for these prediction models in the future.[26–30] In addition, fluorescein angiography, especially taken with ultra-widefield cameras, may provide significant meaningful information for the prediction of post-treatment DME prognoses, and further study utilizing these types of input data is warranted.[31–33] However, solving the problem of structural dissimilarity between different types of input data will still be necessary. The current study is limited by its retrospective nature. A future study comparing real-world data with GAN predictions could improve artificial intelligence integration in clinical practice. In addition, studies comparing GANs with other models, such as convolutional neural networks (CNNs), are also warranted. Regardless, this study evaluated a long-term prediction model for DME using GAN, showing head-to-head comparisons with short-term results, thereby demonstrating the stability and the potential range of application of the model. By visualizing possible future treatment results, GAN models have the advantage of bridging the gap in understanding the disease between clinicians and patients. In addition, the value of this study is highly strengthened by the use of RCT data with long-term regular treatment for DME.

In summary, we demonstrated that GAN models could be effective in predicting the appearance of OCT B-scan images following long-term anti-VEGF treatment in DME eyes. The application of these models may be useful in not only informing treating clinicians, but also in educating patients with DME, regarding the expected clinical course with a tangible and understandable output in the form of an image. This type of information may facilitate patient compliance and aid in management planning.

Acknowledgments

Supported by the National Eye Institute, Bethesda, Maryland (grant #RO1 EY023164), the Department of Ophthalmology at the Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, John P. Hussman Institute of Human Genomics at the University of Miami Miller School of Medicine, Miami, Florida, and the Institute for Computational Biology, Case Western Reserve University, Cleveland, Ohio. This study was partially supported by research grants from the National Institute of Health (R01EY024378 and R01EY034218), and Research to Prevent Blindness/Dr. H. James and Carole Free Catalyst Award for Innovative Research Approaches for AMD. Funds were also received from the F. M. Kirby Foundation.

Institutional Review Board Statement: The study is approved by IRB protocol number 14-001476 of Doheny Eye Institute.

Author Contributions: Study concept and design: S.S. and J.B. Acquisition, analysis, or interpretation of data: J.B., M.A., A.M., M.G.N., G.C., and S.S. Drafting of the manuscript: J.B. Critical revision of the manuscript for important intellectual content and statistical analysis: J.B. and S.S. Study supervision: S.S.

Disclosure: **J. Baek**, None; **Y. He**, None; **M. Emamverdi**, None; **A. Mahmoudi**, None; **M.G. Nittala**, None; **G. Corradetti**, None; **M. Ip**, None; **S.R. Sadda**, Amgen (C), Allergan (C), Regeneron (C), Roche/Genentech (C), Novartis (C), Merck (C), 4DMT (C), Optos (C), Heidelberg (C), Centervue (C), Topcon (F, N), Nidek (F, N), Heidelberg (F, N), Centervue (F, N), Optos (F, N), Carl Zeiss Meditec (F, N), outside the submitted work

References

1. Teo ZL, Tham YC, Yu M, et al. Global prevalence of diabetic retinopathy and projection of burden through 2045: systematic review and meta-analysis. *Ophthalmology*. 2021;128:1580–1591.
2. Bressler NM, Beaulieu WT, Glassman AR, et al. Persistent macular thickening following intravitreous aflibercept, bevacizumab, or ranibizumab for central-involved diabetic macular edema with vision impairment: a secondary analysis of a randomized clinical trial. *JAMA Ophthalmol*. 2018;136:257–269.
3. Bressler NM, Beaulieu WT, Glassman AR, et al. Persistent macular thickening following intravitreous aflibercept, bevacizumab, or ranibizumab for central-involved diabetic macular edema with vision impairment: a secondary analysis of a randomized clinical trial. *JAMA Ophthalmol*. 2018;136:257–269.
4. Baker CW, Glassman AR, Beaulieu WT, et al. Effect of initial management with aflibercept vs laser photocoagulation vs observation on vision loss among patients with diabetic macular edema involving the center of the macula and good visual acuity: a randomized clinical trial. *JAMA*. 2019;321:1880–1894.
5. Brown DM, Schmidt-Erfurth U, Do DV, et al. Intravitreal aflibercept for diabetic macular edema: 100-week results from the VISTA and VIVID studies. *Ophthalmology*. 2015;122:2044–2052.
6. Huang H, Jansonius NM, Chen H, Los LI. Hyperreflective dots on OCT as a predictor of treatment outcome in diabetic macular edema: a systematic review. *Ophthalmol Retina*. 2022;6:814–827.
7. Hui VWK, Szeto SKH, Tang F, et al. Optical coherence tomography classification systems for diabetic macular edema and their associations with visual outcome and treatment responses – an updated review. *Asia Pac J Ophthalmol (Phila)*. 2022;11:247–257.
8. Cao J, You K, Jin K, et al. Prediction of response to anti-vascular endothelial growth factor treatment in diabetic macular oedema using an optical coherence tomography-based machine learning method. *Acta Ophthalmol*. 2021;99:e19–e27.
9. Gallardo M, Munk MR, Kurmann T, et al. Machine learning can predict anti-VEGF treatment demand in a treat-and-extend regimen for patients with neovascular AMD, DME, and RVO associated macular edema. *Ophthalmol Retina*. 2021;5:604–624.
10. Gerendas BS, Bogunovic H, Sadeghipour A, et al. Computational image analysis for prognosis determination in DME. *Vision Res*. 2017;139:204–210.
11. Shi R, Leng X, Wu Y, Zhu S, Cai X, Lu X. Machine learning regression algorithms to predict short-term efficacy after anti-VEGF treatment in diabetic macular edema based on real-world data. *Sci Rep*. 2023;13:18746.

12. Lin TY, Chen HR, Huang HY, et al. Deep learning to infer visual acuity from optical coherence tomography in diabetic macular edema. *Front Med (Lausanne)*. 2022;9:1008950.
13. Chen S-C, Chiu H-W, Chen C-C, Woung L-C, Lo C-M. A novel machine learning algorithm to automatically predict visual outcomes in intravitreal ranibizumab-treated patients with diabetic macular edema. *J Clin Med*. 2018;7:475.
14. Lee H, Kim S, Kim MA, Chung H, Kim HC. Post-treatment prediction of optical coherence tomography using a conditional generative adversarial network in age-related macular degeneration. *Retina*. 2021;41:572–580.
15. Liu S, Hu W, Xu F, et al. Prediction of OCT images of short-term response to anti-VEGF treatment for diabetic macular edema using different generative adversarial networks. *Photodiagnosis Photodyn Ther*. 2023;41:103272.
16. Moon S, Lee Y, Hwang J, et al. Prediction of anti-vascular endothelial growth factor agent-specific treatment outcomes in neovascular age-related macular degeneration using a generative adversarial network. *Sci Rep*. 2023;13:5639.
17. Xu F, Yu X, Gao Y, et al. Predicting OCT images of short-term response to anti-VEGF treatment for retinal vein occlusion using generative adversarial network. *Front Bioeng Biotechnol*. 2022;10:914964.
18. Kong LL, Lian C, Huang D, Li Z, Hu Y, Zhou Q. Breaking the dilemma of medical image-to-image translation. *Thirty-Fifth Conference on Neural Information Processing Systems*. 2021. arXiv preprint. Available at: https://arxiv.org/abs/2110.06465.
19. Teng P-Y. *Caserel - An Open Source Software for Computer-aided Segmentation of Retinal Layers in Optical Coherence Tomography Images*. Geneva, Switzerland: Zenodo; 2013.
20. Ometto G, Moghul I, Montesano G, et al. ReLayer: a free, online tool for extracting retinal thickness from cross-platform OCT images. *Transl Vis Sci Technol*. 2019;8:25.
21. Zhou W, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Proc*. 2004;13:600–612.
22. Coronado I, Pachade S, Trucco E, et al. Synthetic OCT-A blood vessel maps using fundus images and generative adversarial networks. *Sci Rep*. 2023;13:15325.
23. Welander P, Karlsson S, Eklund A. Generative adversarial networks for image-to-image translation on multi-contrast MR images - a comparison of CycleGAN and UNIT. arXiv Preprint. Available at: https://arxiv.org/abs/1806.07777.
24. Staurenghi G, Feltgen N, Arnold JJ, et al. Impact of baseline Diabetic Retinopathy Severity Scale scores on visual outcomes in the VIVID-DME and VISTA-DME studies. *Br J Ophthalmol*. 2018;102:954–958.
25. Xu J, Ishikawa H, Wollstein G, Kagemann L, Schuman JS. Alignment of 3-D optical coherence tomography scans to correct eye movement using a particle filtering. *IEEE Trans Med Imaging*. 2012;31:1337–1345.
26. Paine SK, Bhattacharjee CK, Bhaduri G, et al. Pre-therapeutic biomarkers for ranibizumab therapy among type 2 diabetic patients with diabetic macular edema. *Optom Vis Sci*. 2021;98:81–87.
27. Sorour OA, Levine ES, Baumal CR, et al. Persistent diabetic macular edema: definition, incidence, biomarkers, and treatment methods. *Surv Ophthalmol*. 2023;68:147–174.
28. Custo Greig E, Brigell M, Cao F, et al. Macular and peripapillary optical coherence tomography angiography metrics predict progression in diabetic retinopathy: a sub-analysis of TIME-2b study data. *Am J Ophthalmol*. 2020;219:66–76.
29. Yang D, Tang Z, Ran A, et al. Assessment of parafoveal diabetic macular ischemia on optical coherence tomography angiography images to predict diabetic retinal disease progression and visual acuity deterioration. *JAMA Ophthalmol*. 2023;141:641–649.
30. Wykoff CC, Elman MJ, Regillo CD, Ding B, Lu N, Stoilov I. Predictors of diabetic macular edema treatment frequency with ranibizumab during the open-label extension of the RIDE and RISE trials. *Ophthalmology*. 2016;123:1716–1721.
31. Jiang AC, Srivastava SK, Hu M, et al. Quantitative ultra-widefield angiographic features and associations with diabetic macular edema. *Ophthalmol Retina*. 2020;4:49–56.
32. Fang M, Fan W, Shi Y, et al. Classification of regions of nonperfusion on ultra-widefield fluorescein angiography in patients with diabetic macular edema. *Am J Ophthalmol*. 2019;206:74–81.
33. Allingham MJ, Mukherjee D, Lally EB, et al. A quantitative approach to predict differential effects of anti-VEGF treatment on diffuse and focal leakage in patients with diabetic macular edema: a pilot study. *Transl Vis Sci Technol*. 2017;6:7.