tvst

Special Issue

Development of Deep Learning Models to Predict Best-Corrected Visual Acuity from Optical Coherence Tomography

Michael G. Kawczynski[1], Thomas Bengtsson[1], Jian Dai[1], J. Jill Hopkins[1], Simon S. Gao[1], and Jeffrey R. Willis[1]

[1] Genentech, Inc., South San Francisco, CA, USA

Correspondence: Jeffrey R. Willis, Genentech, Inc., 1 DNA Way, South San Francisco, CA 94080, USA. email: willis.jeffrey@gene.com

Received: September 27, 2019
Accepted: August 3, 2020
Published: September 9, 2020

Keywords: deep learning; ocular imaging; tele-ophthalmology; public health ophthalmology; neovascular age-related macular degeneration

Citation: Kawczynski MG, Bengtsson T, Dai J, Hopkins JJ, Gao SS, Willis JR. Development of deep learning models to predict best-corrected visual acuity from optical coherence tomography. Trans Vis Sci Tech. 2020;9(2):51, https://doi.org/10.1167/tvst.9.2.51

Purpose: To develop deep learning (DL) models to predict best-corrected visual acuity (BCVA) from optical coherence tomography (OCT) images from patients with neovascular age-related macular degeneration (nAMD).

Methods: Retrospective analysis of OCT images and associated BCVA measurements from the phase 3 HARBOR trial (NCT00891735). DL regression models were developed to predict BCVA at the concurrent visit and 12 months from baseline using OCT images. Binary classification models were developed to predict BCVA of Snellen equivalent of <20/40, <20/60, and ≤20/200 at the concurrent visit and 12 months from baseline.

Results: The regression model to predict BCVA at the concurrent visit had $R^2 = 0.67$ (root-mean-square error [RMSE] = 8.60) in study eyes and $R^2 = 0.84$ (RMSE = 9.01) in fellow eyes. The best classification model to predict BCVA at the concurrent visit had an area under the receiver operating characteristic curve (AUC) of 0.92 in study eyes and 0.98 in fellow eyes. The regression model to predict BCVA at month 12 using baseline OCT had $R^2 = 0.33$ (RMSE = 14.16) in study eyes and $R^2 = 0.75$ (RMSE = 11.27) in fellow eyes. The best classification model to predict BCVA at month 12 had AUC = 0.84 in study eyes and AUC = 0.96 in fellow eyes.

Conclusions: DL shows promise in predicting BCVA from OCTs in nAMD. Further research should elucidate the utility of models in clinical settings.

Translational Relevance: DL models predicting BCVA could be used to enhance understanding of structure–function relationships and develop more efficient clinical trials.

Introduction

Retinal diseases such as neovascular age-related macular degeneration (nAMD) are characterized by pathophysiological and anatomical changes that can interfere with vision and lead to permanent vision loss. Normalization of the retinal anatomy with effective treatment, such as intravitreal anti-vascular endothelial growth factor injections, can lead to improvement in best-corrected visual acuity (BCVA).[1–3] Although it is generally accepted that retinal anatomy determines visual function in retinal disease, the reported correlation between optical coherence tomography (OCT) and BCVA has been low, and it has proven difficult to determine the precise relationship between structure and function sufficiently to be able to predict visual outcomes from anatomical changes detected in retinal images generated by OCT.[3–9] One solution for improving the accuracy of this relationship is the development of deep learning (DL) models that predict BCVA from macular OCTs rather than relying on conventional correlation analyses.

Conventional correlation analyses are limited in their ability to detect novel relationships between anatomic and visual parameters by the need to identify and prespecify a candidate set of features for analysis,[10–15] a process that is limited by the insight of the human investigator. Moreover, conventional methods that prespecify features, such as central subfield thickness and central foveal thickness, typically yield aggregated measures of retinal health that may not have a sufficiently specific relationship to vision outcomes. For example, in the HARBOR trial data, features derived from intraretinal fluid and total retinal thickness correlate with baseline BCVA at $R^2 = 0.21$.[14] DL algorithms and, in particular, deep convolutional neural networks (CNNs), are free of the limitation of prespecifying features. By evaluating an entire image, CNNs overcome the need for prespecified candidate features by automatically generating and learning among millions of possible (anatomical) structures that may be predictive of the outcome of interest.[16,17] Consequently, DL may provide a possible solution for improving the ability to predict and/or classify visual function from OCT images of retinal anatomy, assuming that there is no other nonretinal ocular pathology. To date, there have been limited studies using DL to quantify the relationship between BCVA and retinal anatomy. A major reason for the lack of studies is the difficulty in obtaining high-quality BCVA data that can be correlated to standardized OCT images. In order to address this limitation, we utilized data from the phase 3 HARBOR clinical trial in nAMD (NCT00891735) that coupled monthly OCT data with Early Treatment Diabetic Retinopathy Study (ETDRS) BCVA measurements over 24 months.

Large randomized clinical trials designed to evaluate new therapies in retinal disease provide a rich source of OCT images and visual outcome measurements suitable for training and evaluation of DL algorithms. In the phase 3 HARBOR clinical trial of the anti-vascular endothelial growth factor agent ranibizumab in patients with nAMD, patients were evaluated monthly for changes in BCVA using standard ETDRS protocols and changes in retinal thickness using OCT. The primary objective of the study presented here was to assess whether DL can automatically predict concurrent and future BCVA from OCT images from patients with nAMD in the HARBOR trial. Specifically, we present DL results related to (1) models that assess the quality of a DL regression model to predict exact BCVA value from OCTs, and (2) models that predict BCVA of <69 letters (Snellen equivalent, 20/40), <59 letters (Snellen equivalent, 20/60), or ≤38 letters (Snellen equivalent, 20/200) from OCTs. For all BCVA outcomes, DL models were evaluated for their ability to predict BCVA from an OCT taken at the same (concurrent) visit and for their ability to predict month 12 BCVA from baseline OCT. Snellen equivalents of 20/40 and 20/60 were chosen because visual acuity worse than these levels is considered to reflect visual acuity impairment based on definitions by the United States and the World Health Organization, respectively.[18] A Snellen equivalent of 20/200 or worse was used to reflect the definition of legal blindness in the United States.[19]

Methods

Source of Dataset

Prospectively collected BCVA measurements and OCT images taken of 1071 patients from the phase 3 HARBOR clinical trial (NCT00891735) were used. HARBOR adhered to the tenets of the Declaration of Helsinki and was compliant with the Health Insurance Portability and Accountability Act. The protocol was approved by each institutional review board before the study began, and all patients provided written informed consent for future medical research and analyses based on results of the trial.

The study design and results of HARBOR have been published previously.[2,20] In summary, 1097 adult patients with treatment-naïve subfoveal choroidal neovascularization (CNV) secondary to nAMD were enrolled if they had BCVA between 20/40 and 20/320 (Snellen equivalent) using standard ETDRS charts and protocols. Patients (one study eye each) were randomized 1:1:1:1 to ranibizumab given according to one of the following treatment regimens: 0.5 mg monthly, 0.5 mg as needed (PRN), 2.0 mg monthly, and 2.0 mg PRN. Patients in the PRN groups received three monthly injections followed by monthly evaluations, with re-treatment only if there was any sign of disease activity on OCT or if there was a ≥5-letter decrease in BCVA from the previous visit.[2] BCVA measurements and OCT images were obtained at baseline and at monthly intervals for 24 months.[20]

OCT Images

Images were collected using spectral-domain Cirrus HD-OCT (Carl Zeiss Meditec, Dublin, CA, USA).[2] Resolution was 200 × 200 × 1024 voxels with a size of $30.0 \times 30.0 \times 2.0\ \mu m^3$, covering a volume of $6 \times 6 \times 2\ mm^3$. The dataset consisted of 50,275 OCT scans from 1071 patients. For each of the 50,275 OCT scans, the retina was flattened to the retinal pigment epithelium layer segmentation provided by Zeiss software, and the volumes were cropped to 384 pixels above and 128 pixels below the flattened retinal pigment epithelium. Thirty slices of 512 × 200 pixels were generated per scan by rotating about the z-axis at angles of 0, 30, 60, 90, 120, and 150 degrees in reference to the center of the cropped volume, offset at –8, –4, 0, 4, and 8 pixels for each of the six angles, resulting in a total of 1,508,250 slices. The OCT dataset was split at

Table 1. Characteristics of the Internal Validation Test Set Used to Evaluate the Models to Predict BCVA from OCT

Characteristic	Study Eyes	Fellow Eyes	All Eyes
Patients with OCT, *n*	147	147	147
OCTs with BCVA, *n*	3616	3610	7226
Images, *n*	108,480	108,300	216,780
BCVA, mean (SD)			
Baseline	53.93 (13.20)	69.46 (22.92)	61.66 (20.20)
Month 6	64.74 (15.05)	70.60 (22.55)	67.65 (19.33)
Month 12	63.87 (16.96)	69.49 (23.15)	66.69 (20.46)
Month 18	63.19 (17.81)	69.98 (21.73)	66.58 (20.12)
Month 24	65.02 (17.12)	68.56 (22.56)	66.78 (20.06)
All time points[a]	63.24 (15.98)	69.88 (22.40)	66.56 (19.73)
Mean of all visits	62.68 (14.92)	69.49 (22.42)	66.08 (19.31)
BCVA range (min, max)			
Baseline	55 (19, 74)	93 (0, 93)	93 (0, 93)
Month 6	83 (10, 93)	98 (0, 98)	98 (0, 98)
Month 12	89 (0, 89)	100 (0, 100)	100 (0, 100)
Month 18	82 (8, 90)	96 (0, 96)	96 (0, 96)
Month 24	78 (12, 90)	98 (0, 98)	98 (0, 98)
All time points[a]	93 (0, 93)	100 (0, 100)	100 (0, 100)
Mean of all visits	72.97 (13.23, 86.20)	95.40 (0.04, 95.44)	95.40 (0.04, 95.44)
Patients with OCT at baseline and BCVA at month 12, *n*	126	125	126
Baseline images from patients with baseline OCT and BCVA at month 12, *n*	3780	3750	7530

[a] BCVA from screening through month 24.

the patient level into (1) a randomly selected internal validation test set of 147 patients to be used for evaluation (Table 1), and (2) a set of 924 patients that was further split into five folds to be used for model development via cross-validation (Table 2). The patients in each fold remained constant for each outcome variable.

Outcome Variables for DL Modeling

BCVA

The BCVA outcomes of interest were (1) BCVA in ETDRS letters at each study visit, and (2) whether a specific BCVA value was <69 letters (Snellen equivalent, 20/40), <59 letters (Snellen equivalent, 20/60), or ≤38 letters (Snellen equivalent, 20/200). Snellen equivalents of 20/40 and 20/60 were chosen because they are considered to reflect functionally meaningful levels of visual acuity impairment,[18] and a Snellen equivalent of 20/200 or worse was used to reflect the definition of legal blindness in the United States.[19]

The mean (±SD) BCVA of the study eyes in the internal validation test set was 53.93 (±13.20) letters at baseline and 65.02 (±17.12) letters at month 24; the mean (±SD) BCVA of the fellow eyes was 69.46 (±22.92) letters at baseline and 68.56 (±22.56) letters at month 24 (Table 1). Visual acuity in the study eyes at baseline, month 6, month 12, month 18, and month 24 was 55, 83, 89, 82, and 78 letters, respectively (Table 1). Visual acuity in the fellow eyes at baseline, month 6, month 12, month 18, and month 24 was 93, 98, 100, 96, and 98 letters, respectively (Table 1).

DL Algorithms

DL models were evaluated for their ability to predict (1) exact BCVA value in ETDRS letters from OCT images obtained on the same visit; (2) exact BCVA at month 12 from baseline OCT; (3) BCVA of <69 letters (Snellen equivalent, 20/40), <59 letters (Snellen equivalent, 20/60), or ≤38 letters (Snellen equivalent, 20/200) from OCT images obtained on the same visit; and (4)

Table 2. Characteristics of the OCT Dataset to Develop the DL Models to Predict BCVA from OCT

Characteristic	Study Eyes	Fellow Eyes	All Eyes
Patients with OCT, *n*	924	919	924
OCTs with BCVA, *n*	21,623	21,426	43,049
Images, *n*	648,690	642,780	1,291,470
BCVA, mean (SD)			
Baseline	54.18 (12.73)	67.22 (23.33)	60.65 (19.85)
Month 6	62.89 (15.92)	70.26 (21.30)	66.55 (19.14)
Month 12	63.92 (16.87)	69.72 (22.30)	66.81 (19.97)
Month 18	63.49 (17.31)	69.85 (21.68)	66.66 (19.86)
Month 24	63.12 (17.97)	68.93 (22.18)	66.01 (20.38)
All time points[a]	62.39 (16.29)	69.49 (22.01)	65.92 (19.68)
Mean of all visits	61.68 (15.00)	68.34 (22.35)	65.00 (19.31)
BCVA range (min, max)			
Baseline	75 (3, 78)	97 (0, 97)	97 (0, 97)
Month 6	88 (6, 94)	99 (0, 99)	99 (0, 99)
Month 12	93 (2, 95)	100 (0, 100)	100 (0, 100)
Month 18	89 (6, 95)	100 (0, 100)	100 (0, 100)
Month 24	96 (0, 96)	100 (0, 100)	100 (0, 100)
All time points[a]	99 (0, 99)	100 (0, 100)	100 (0, 100)
Mean of all visits	81.03 (9.93, 90.96)	97.59 (0, 97.59)	97.59 (0, 97.59)
Patients with OCT at baseline and BCVA at month 12, *n*	720	708	722
Baseline images from patients with baseline OCT and BCVA at month 12, *n*	21,600	21,240	42,840

[a] BCVA from screening through month 24.

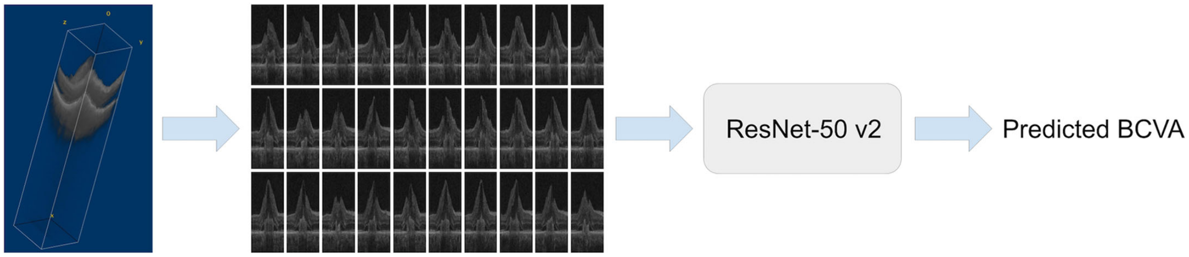

Figure 1. DL pipeline. Predicting BCVA by using three-dimensional OCT volume represented as 30 two-dimensional images input to a ResNet-50 v2 CNN.

BCVA of <69, <59, or ≤38 letters at month 12 from baseline OCT.

Predicting BCVA at the Concurrent Visit

All DL modeling was performed using TensorFlow 1.14.0 with Keras 2.2.5 on a NVIDIA V100 GPU (NVIDIA, Santa Clara, CA, USA) and the ResNet-50 v2 CNN architecture.[21,22] Individual slices of 512 × 200 pixels were randomly shuffled from the training set and fed into the CNN using a batch size of 64 images. The model was trained to predict BCVA at the same visit as the OCT scan (Fig. 1). For regression, layers of global average pooling and dropout (0.85) were added using L_2 regularization (0.05) on the final dense layer with a linear activation function.[23,24] The loss function was mean squared error, and the optimizer was Rectified Adam (RAdam).[25] The model was trained for only one epoch for each cross-validation fold. For classification, the architecture remained the

Table 3. Performance of the DL Model for Regression of BCVA on OCT

	Regression of BCVA on OCT											
	Study Eyes				Fellow Eyes				All Eyes			
	R^2	95% CI	RMSE	No.	R^2	95% CI	RMSE	No.	R^2	95% CI	RMSE	No.
Baseline	0.24	0.12, 0.37	11.55	126	0.80	0.72, 0.85	10.35	125	0.66	0.59, 0.73	11.75	251
Month 6	0.54	0.42, 0.65	10.22	136	0.80	0.73, 0.85	10.14	134	0.72	0.66, 0.78	10.18	270
Month 12	0.62	0.51, 0.71	10.46	142	0.82	0.75, 0.86	9.95	143	0.75	0.70, 0.80	10.18	285
Month 18	0.63	0.52, 0.72	10.85	138	0.80	0.74, 0.86	9.67	138	0.74	0.68, 0.79	10.25	276
Month 24	0.62	0.51, 0.72	10.54	133	0.79	0.72, 0.85	10.38	132	0.73	0.67, 0.78	10.42	265
All time points	0.55	0.53, 0.57	10.70	3616	0.79	0.77, 0.80	10.37	3610	0.71	0.70, 0.73	10.55	7226
Mean of all visits	0.67	0.57, 0.75	8.60	147	0.84	0.78, 0.88	9.01	147	0.79	0.75, 0.83	8.78	294

CI, confidence interval.

same, except the final layer used a softmax activation function with a sparse categorical cross-entropy loss function. Models were initialized with the weights from the regression model for each fold. Models were trained for two epochs with the base model layers untrainable using the Adam optimizer, then an additional one epoch with the base model layers trainable using a stochastic gradient descent (SGD) optimizer.

Predicting BCVA at 12 Months from Baseline

For the regression task predicting BCVA at month 12 from baseline OCT images, the architecture was the same as the regression architecture described above, except dropout was increased to 0.995. For each fold, models were initialized with the weights from the regression model trained to predict BCVA at the concurrent visit. The first three epochs were trained using the SGD optimizer with the base model layers untrainable and then were trained for an additional 1000 epochs with the base model layers trainable using the SGD optimizer. For classification, the architecture was the same as the classification architecture described above. For each fold, models were initialized with the weights from the regression model trained to predict BCVA at 12 months from baseline. Models were trained for 20 epochs with the base model layers untrainable using the RAdam optimizer. The weights used to predict were chosen from the epoch with the lowest validation loss in each fold.

Evaluation of the DL Models

The metrics to evaluate the model fits at a particular visit were calculated at the eye level by the average of the predictions per eye generated for the 30 slices from each of the five development models on the out-of-sample internal validation test set. In other words, each of the five cross-validation models that saw 80% of the data from the training set were used to generate a prediction for each of the 30 slices per eye in the test set, resulting in 150 predictions per eye, of which the mean of the 150 predictions was taken. Furthermore, to evaluate model performance at the concurrent visit across all of the visits while accounting for the potential bias of repeated measurements of the same eye, we took the mean of all visits for the regression task and randomly selected a visit for each patient in the classification task. The R^2 value, root-mean-square error (RMSE), and mean difference (MD) and 95% limits of agreement (LOA) from Bland–Altman plots were used to evaluate the DL regression models, whereas the area under the receiver operating characteristic curve (AUC) and the area under the precision-recall curve (AUPRC) were used to assess the performance of the DL models for classification. To understand if the DL prediction of month 12 BCVA from baseline OCT contributed additional information compared with using baseline BCVA alone, linear models were fit using the R statistical programming language (R Foundation for Statistical Computing, Vienna, Austria) to predict BCVA at month 12 from (1) a univariable input of the DL prediction of month 12 BCVA from baseline OCT, (2) a univariable input of baseline BCVA, and (3) a multivariable input of both the DL prediction of month 12 BCVA from baseline OCT and baseline BCVA. Additionally, results from the five-fold cross-validation tuning set are reported using the mean of the 30 tuning predictions per eye per visit (Supplementary Tables S1–S4).

Results

Predicting BCVA at the Concurrent Visit

Regression Results

In the study eyes, the DL model to predict BCVA at the concurrent visit had $R^2 = 0.24$, RMSE $= 11.55$,

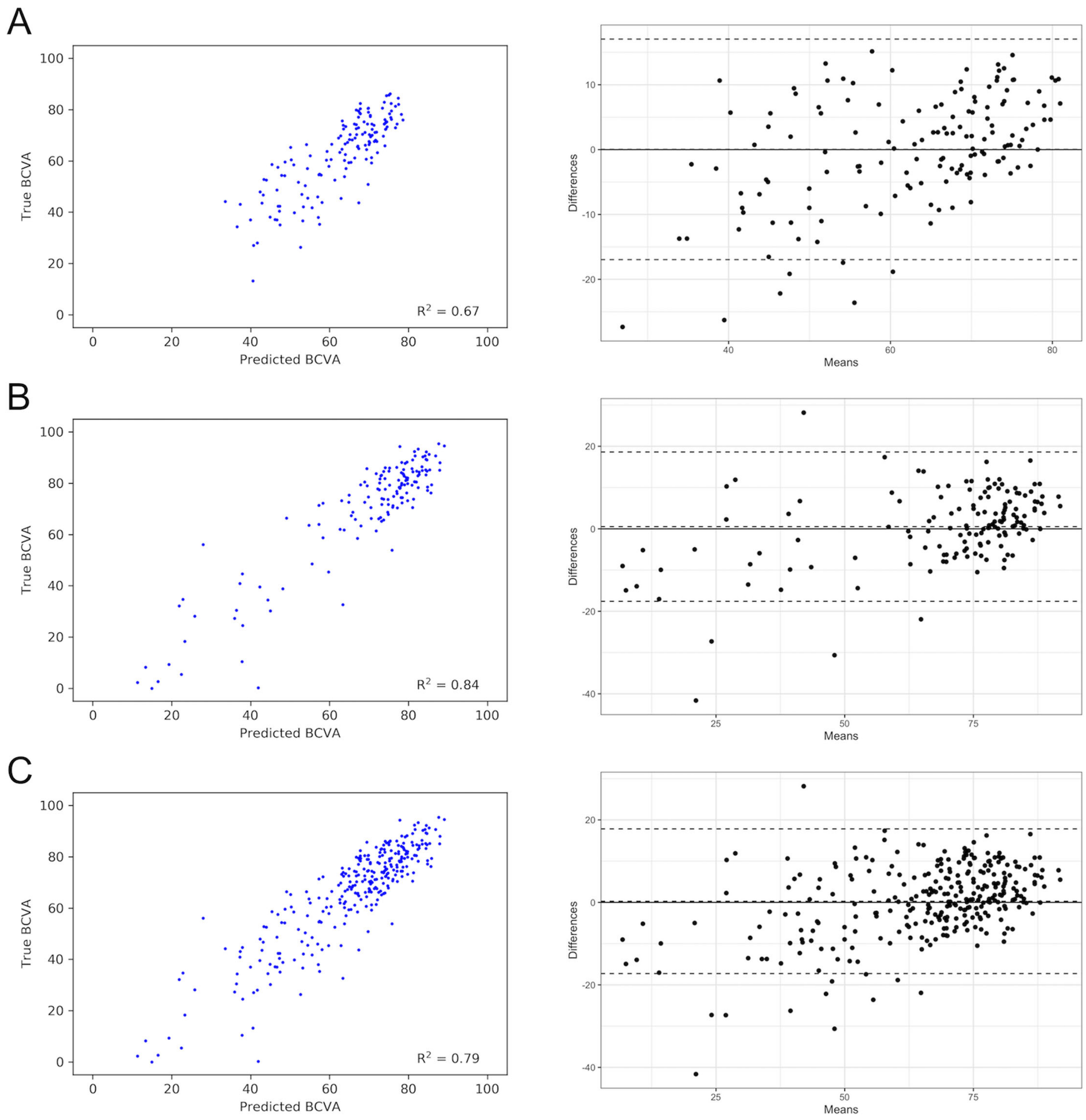

Figure 2. Actual versus predicted BCVA at the concurrent visit and performance of DL algorithms that analyze OCT images to predict BCVA for the concurrent visit. (A) Study eye mean over all visits: $R^2 = 0.67$, RMSE = 8.60, MD = 0.04 letters, 95% LOA = −16.96 to 17.04 letters. (B) Fellow eye mean over at all visits: $R^2 = 0.84$, RMSE = 9.01, MD = 0.51 letters, 95% LOA = −17.58 to 18.61 letters. (C) Both study and fellow eye mean over all visits: $R^2 = 0.79$, RMSE = 8.78, MD = 0.28 letters, 95% LOA = −17.26 to 17.81 letters.

MD = –1.81 letters, and 95% LOA = −26.57 to 22.95 letters at baseline, and $R^2 = 0.67$, RMSE = 8.60, MD = 0.04 letters, and 95% LOA = −16.96 to 17.04 letters for the mean over all of the visits (Table 3; Fig. 2A). In the fellow eyes, $R^2 = 0.80$, RMSE = 10.35, MD = –1.86 letters, and 95% LOA = −23.03 to 19.31 letters at baseline, and $R^2 = 0.84$, RMSE = 9.01, MD = 0.51 letters, and 95% LOA = −17.58 to 18.61 letters for the mean over all of the visits (Table 3; Fig. 2B). In all eyes, $R^2 = 0.66$, RMSE = 11.75, MD = –1.84 letters, and 95% LOA = −24.83 to 21.16 letters at baseline, and $R^2 = 0.79$, RMSE = 8.78, MD = 0.28 letters, and 95% LOA = −17.26 to 17.81 letters for the mean over all of the visits (Table 3; Fig. 2C).

Binary Classifications

BCVA of <69 letters (Snellen equivalent, 20/40). In the study eyes, the DL model to predict BCVA of <69 letters (Snellen equivalent, 20/40) had AUC = 0.89

Table 4. Performance of the DL Models for Binary Classification of BCVA of <69 Letters (Snellen Equivalent, 20/40), <59 Letters (Snellen Equivalent, 20/60), and ≤38 Letters (Snellen Equivalent, 20/200) at the Concurrent Visit from Associated OCT

Predicted BCVA from	Study Eyes			Fellow Eyes			All Eyes		
Concurrent OCT	AUC	95% CI	No.	AUC	95% CI	No.	AUC	95% CI	No.
<69 letters	0.89	0.83, 0.94	147	0.93	0.89, 0.98	147	0.92	0.89, 0.95	294
<59 letters	0.92	0.87, 0.97	147	0.97	0.94, 1.00	147	0.95	0.92, 0.97	294
≤38 letters	0.92	0.87, 0.96	147	0.98	0.96, 1.00	147	0.96	0.93, 0.98	294

CI, confidence interval.

for one concurrent visit randomly taken per eye and AUPRC = 0.88 with a class balance of 72 positive eyes and 75 negative eyes (Table 4; Fig. 3A). In the fellow eyes, the DL model to predict BCVA of <69 letters (Snellen equivalent, 20/40) had AUC = 0.93 for one concurrent visit randomly taken per eye and AUPRC = 0.97 with a class balance of 103 positive eyes and 44 negative eyes (Table 4; Fig. 3B). In all eyes, the DL model to predict BCVA of <69 letters (Snellen equivalent, 20/40) had AUC = 0.92 for one concurrent visit randomly taken per eye and AUPRC = 0.94 with a class balance of 175 positive eyes and 119 negative eyes (Table 4; Fig. 3C).

BCVA of <59 letters (Snellen equivalent, 20/60). In the study eyes, the DL model to predict BCVA of <59 letters had AUC = 0.92 for one concurrent visit randomly taken per eye and AUPRC = 0.95 with a class balance of 100 positive eyes and 47 negative eyes (Table 4). In the fellow eyes, the DL model to predict BCVA of <59 letters had AUC = 0.97 for one concurrent visit randomly taken per eye and AUPRC = 0.99 with a class balance of 114 positive eyes and 33 negative eyes (Table 4). In all eyes, the DL model to predict BCVA of <59 letters had AUC = 0.95 for one concurrent visit randomly taken per eye and AUPRC = 0.98 with a class balance of 214 positive eyes and 80 negative eyes (Table 4).

BCVA of ≤38 letters (Snellen equivalent, 20/200). In the study eyes, the DL model to predict BCVA of ≤38 letters had AUC = 0.92 for one concurrent visit randomly taken per eye and AUPRC = 0.99 with a class balance of 113 positive eyes and 14 negative eyes (Table 4). In the fellow eyes, the DL model to predict BCVA of ≤38 letters had AUC = 0.98 for one concurrent visit randomly taken per eye and AUPRC = 1.00 with a class balance of 129 positive eyes and 18 negative eyes (Table 4). In all eyes, the DL model to predict BCVA of ≤38 letters had AUC = 0.96 for one concurrent visit randomly taken per eye and AUPRC = 0.99 with a class balance of 262 positive eyes and 32 negative eyes (Table 4).

Predicting BCVA at 12 Months from Baseline OCT

Regression Results

The characteristics of the dataset used to evaluate the ability of the model to predict month 12 BCVA from baseline OCT are shown in Table 1. The DL model to predict BCVA at 12 months from baseline OCT had $R^2 = 0.33$, 0.75, and 0.58 and RMSE = 14.16, 11.27, and 13.25 for study eyes, fellow eyes, and all eyes, respectively (Table 5; Figs. 4A–4C). The DL model to predict BCVA at 12 months from baseline OCT had MD = –1.63 letters and 95% LOA = −29.48 to 26.22 letters for study eyes, MD = –2.31 letters and 95% LOA = −29.96 to 25.33 letters for fellow eyes, and MD = –1.97 letters and 95% LOA = −29.67 to 25.73 letters for all eyes (Figs. 4A–4C). The multivariable linear model to predict BCVA at 12 months from both DL predictions of month 12 BCVA from baseline OCT and baseline BCVA had $R^2 = 0.40$, 0.88, and 0.68 for study eyes, fellow eyes, and all eyes, respectively (Table 5).

Binary Classifications

BCVA of <69 letters (Snellen equivalent, 20/40). The DL model to predict month 12 BCVA of <69 letters from baseline OCT had AUC = 0.80, 0.92, and 0.87 for study eyes, fellow eyes, and all eyes, respectively (Table 6; Figs. 5A–5C). In study eyes, the DL model to predict month 12 BCVA of <69 letters from baseline OCT had AUPRC = 0.74 with a class balance of 58 positive eyes and 68 negative eyes. In fellow eyes, the DL model to predict month 12 BCVA of <69 letters from baseline OCT had AUPRC = 0.97 with a class balance of 88 positive eyes and 37 negative eyes. In all eyes, the DL model to predict month 12 BCVA of <69 letters from baseline OCT had AUPRC = 0.91 with a class balance of 146 positive eyes and 105 negative eyes.

BCVA of <59 letters (Snellen equivalent, 20/60). The DL model to predict month 12 BCVA of <59 letters from baseline OCT had AUC = 0.84, 0.93, and 0.89 for study eyes, fellow eyes, and all eyes,

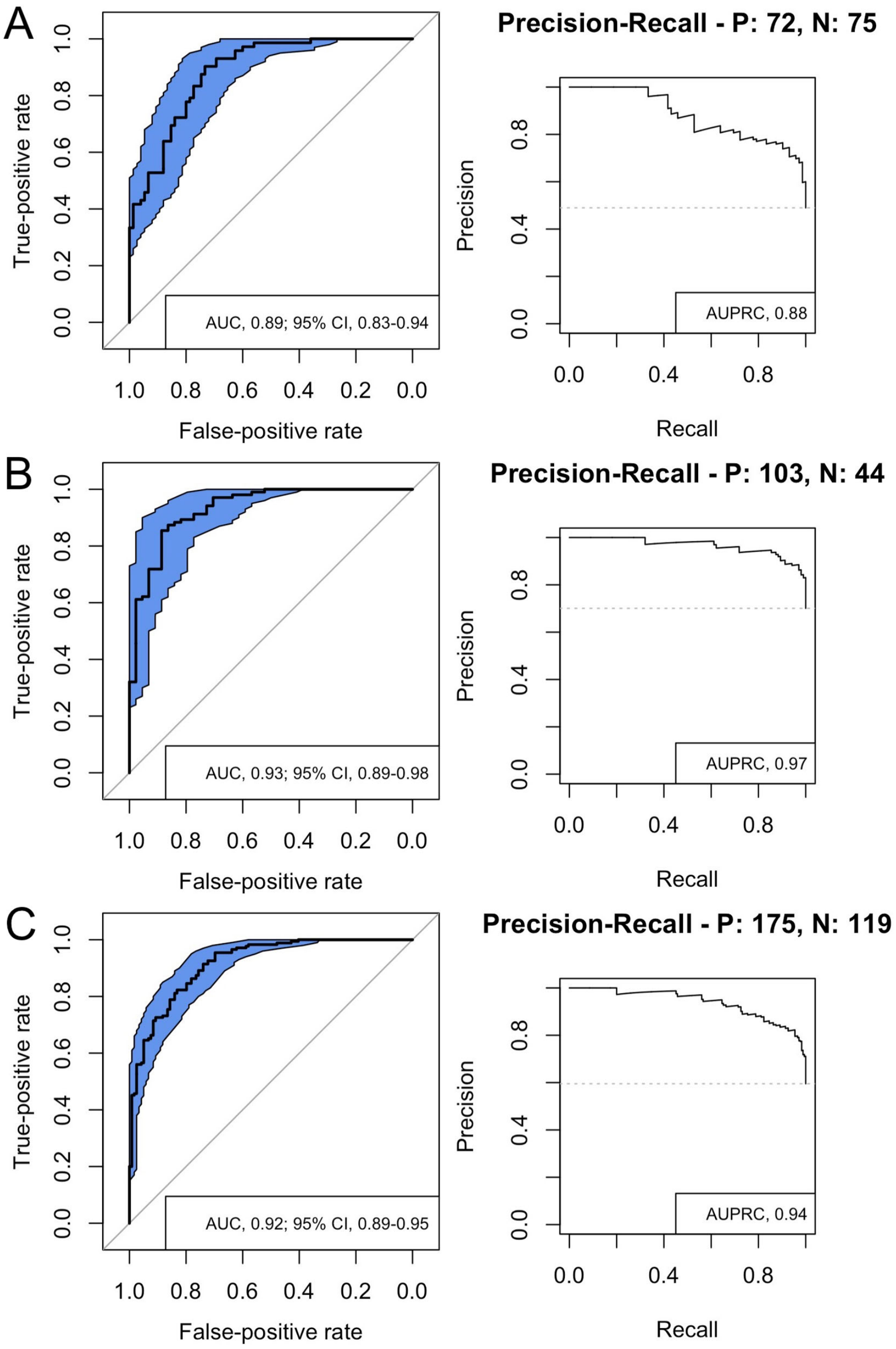

Figure 3. Performance of DL algorithms that predict BCVA of <69 letters at the concurrent visit from associated OCT. (A) BCVA of <69 letters, study eye at random visit: AUC = 0.89, AUPRC = 0.88. (B) BCVA of <69 letters, fellow eye at random visit: AUC = 0.93, AUPRC = 0.97. (C) BCVA of <69 letters, both study and fellow eye at random visit: AUC = 0.92, AUPRC = 0.94. CI, confidence interval; N, negative cases (<69 letters); P, positive cases (≥69 letters).

respectively (Table 6). In study eyes, the DL model to predict month 12 BCVA of <59 letters from baseline OCT had AUPRC = 0.90 with a class balance of 83 positive eyes and 43 negative eyes. In fellow eyes, the DL model to predict month 12 BCVA of <59 letters from baseline OCT had AUPRC = 0.98 with a class balance of 101 positive eyes and 24 negative eyes. In all eyes, the DL model to predict month 12 BCVA of <59 letters from baseline OCT had AUPRC = 0.95 with a class balance of 184 positive eyes and 67 negative eyes.

BCVA of ≤38 letters (Snellen equivalent, 20/200). The DL model to predict month 12 BCVA of ≤38 letters from baseline OCT had AUC = 0.77, 0.96, and

Table 5. Performance of Linear Model for Regression of BCVA at Month 12 from Baseline OCT and Baseline Letters

	Linear Regression of Month 12 BCVA from Baseline OCT and Baseline BCVA											
	Study Eyes				Fellow Eyes				All Eyes			
	R^2	95% CI	RMSE	No.	R^2	95% CI	RMSE	No.	R^2	95% CI	RMSE	No.
Baseline OCT	0.33[a]	0.20, 0.47	14.16	126	0.75[a]	0.62, 0.82	11.27	125	0.58[a]	0.49, 0.65	13.25	251
Baseline BCVA	0.25[a]	0.13, 0.39	14.98	126	0.87[a]	0.82, 0.91	8.04	125	0.62[a]	0.54, 0.69	12.49	251
Baseline OCT + baseline BCVA	0.40[b]	0.24, 0.54	13.45	126	0.88[b]	0.81, 0.93	7.70	125	0.68[b]	0.57, 0.75	11.61	251

CI, confidence interval.

[a] $P < 0.001$ for univariable model compared with the null hypothesis.

[b] $P < 0.001$ for each coefficient in the multivariable model.

Table 6. Performance of the DL Models for Binary Classification of Month 12 BCVA of <69 Letters (Snellen Equivalent, 20/40), <59 Letters (Snellen Equivalent, 20/60), and ≤38 Letters (Snellen Equivalent, 20/200) from Baseline OCT

Predicted BCVA at Month 12 from Baseline OCT	Study Eyes			Fellow Eyes			All Eyes		
	AUC	95% CI	No.	AUC	95% CI	No.	AUC	95% CI	No.
<69 letters	0.80	0.72, 0.87	126	0.92	0.88, 0.97	125	0.87	0.83, 0.91	251
<59 letters	0.84	0.76, 0.92	126	0.93	0.88, 0.99	125	0.89	0.85, 0.94	251
≤38 letters	0.77	0.66, 0.88	126	0.96	0.92, 1.00	125	0.89	0.84, 0.95	251

CI, confidence interval.

0.89 for study eyes, fellow eyes, and all eyes, respectively (Table 6). In study eyes, the DL model to predict month 12 BCVA of ≤38 letters from baseline OCT had AUPRC = 0.97 with a class balance of 114 positive eyes and 12 negative eyes. In fellow eyes, the DL model to predict month 12 BCVA of ≤38 letters from baseline OCT had AUPRC = 0.99 with a class balance of 109 positive eyes and 16 negative eyes. In all eyes, the DL model to predict month 12 BCVA of ≤38 letters from baseline OCT had AUPRC = 0.98 with a class balance of 223 positive eyes and 28 negative eyes.

Discussion

As shown by our results, DL models show promise for predicting BCVA from OCT images in patients with nAMD. The predictive accuracy of the derived model was greatest in the fellow eyes, reaching a correlation between mean predicted and mean observed BCVA of ~0.92 ($R^2 = 0.84$, RMSE = 9.01) (Table 3). A moderate to strong correlation was also seen for BCVA outcomes in the study eyes, with correlations of ~0.49 ($R^2 = 0.24$, RMSE = 11.55) at baseline (before treatment) and ~0.79 ($R^2 = 0.62$, RMSE = 10.54) at month 24 (after treatment) (Table 3).

To benchmark the presented model results (RMSE of ~10-letter error), we have generated predictions based on the screening BCVA to the baseline BCVA in HARBOR using simple linear regression (Supplementary Fig. S1). The average number of days between screening and baseline visits was 8.3 days, with a mean change of 0.6 letters; for this BCVA-to-BCVA prediction, the RMSE (from the regression) is 6.1 letters. This average error of 6.1 letters likely represents an information limit inherent to BCVA observations in HARBOR. Previous work on intersession repeatability of visual acuity scores in nAMD has reported an error of ~12 letters.[26] We also note that the average BCVA improvement to anti-vascular endothelial growth factor treatment in HARBOR was reported to be 7.6 to 9.1 letters. The quality of our DL prediction model, which makes predictions from OCT to BCVA, should be compared relative to the above (information) limits.

The results suggest the existence of a mapping (defined by the DL models) between retinal structures and visual function in nAMD. Consequently, OCT images could provide a means of indirectly measuring visual function in clinical trials, as well as in evolving clinical practice settings, such as telemedicine or home monitoring. Further research is required to elucidate the utility of DL models in clinical research and clinical care.

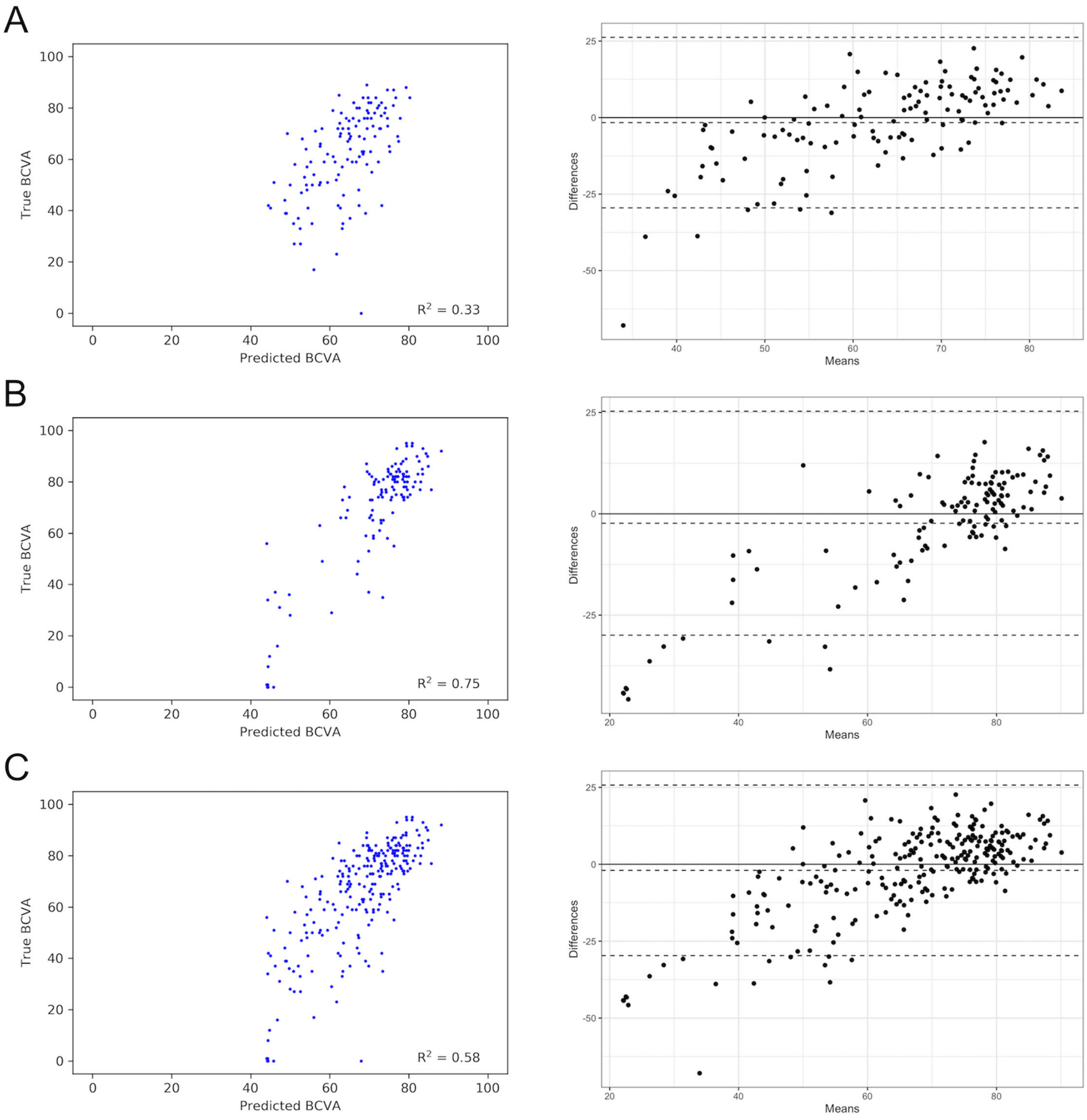

Figure 4. Actual versus predicted BCVA at month 12. Performance of DL algorithms that analyze baseline OCT images to predict BCVA at month 12. (A) Study eyes: $R^2 = 0.33$, RMSE = 14.16, MD = –1.63 letters, 95% LOA = −29.48 to 26.22 letters. (B) Fellow eyes: $R^2 = 0.75$, RMSE = 11.27, MD = –2.31 letters, 95% LOA = −29.96 to 25.33 letters. (C) Both study and fellow eyes: $R^2 = 0.58$, RMSE = 13.25, MD = –1.97 letters, 95% LOA = −29.67 to 25.73 letters.

The difference in model performance between predicting visual function in the study eyes versus the fellow eyes was expected due to the restricted range in BCVA at baseline in the study eyes. Specifically, the eligibility criteria for HARBOR required that the study eyes have some vision loss and subfoveal CNV at baseline, with BCVA between 20/40 and 20/320 (Snellen equivalent).[2] These criteria were not required for the fellow eyes. Hence, the restricted range of BCVA in the study eyes at baseline (SD = 13.2) (Table 1) reduced the dynamic range and led to a more challenging regression task compared with the task of predicting BCVA in the fellow eyes, which had greater variability in BCVA at baseline (baseline SD = 22.9) (Table 1). This observation is supported by the fact that the predictive accuracy of concurrent prediction of BCVA from OCT increases in the study eyes over the course of the trial, along with an increase in the variability of BCVA (Tables 1–3; Supplementary Table S1). This increase in range after treatment is consistent with an

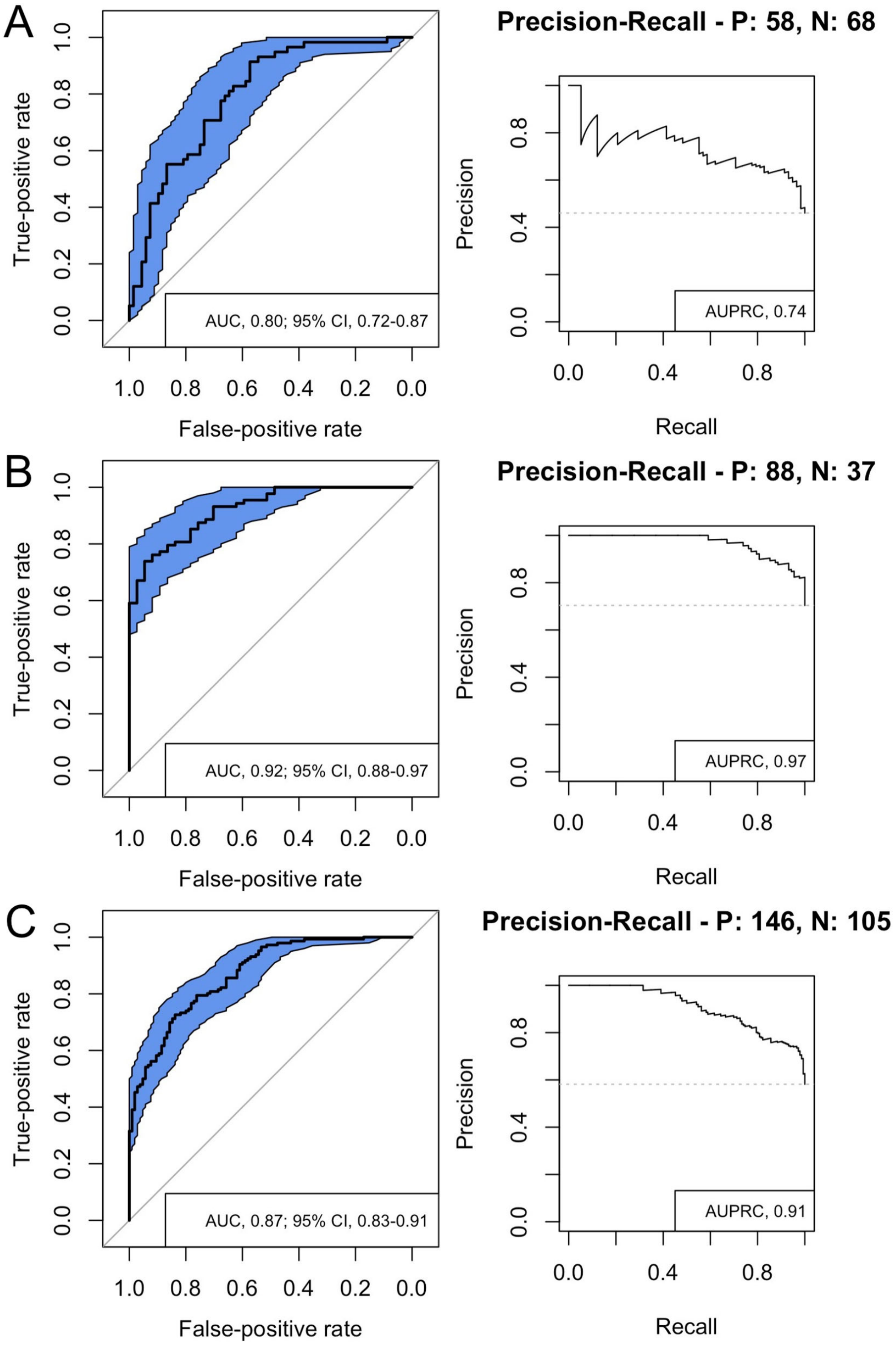

Figure 5. Performance of DL algorithms that predict BCVA of <69 letters at month 12 from baseline OCT. (A) Month 12 BCVA of <69 letters, study eyes: AUC = 0.80, AUPRC = 0.74. (B) Month 12 BCVA of <69 letters, fellow eyes: AUC = 0.92, AUPRC = 0.97. (C) Month 12 BCVA of <69 letters, both study and fellow eyes: AUC = 0.87, AUPRC = 0.91. CI, confidence interval; N, negative cases (<69 letters); P, positive cases (≥69 letters).

effective treatment producing visual improvements in many patients. If, by simulation, we (artificially) restrict the variance of BCVA in the fellow eyes to that of the study eyes at baseline (i.e., SD = 13.2), R^2 decreases from 0.80 to 0.33 (Supplementary Fig. S2). Similarly, if we plot log(var(BCVA)) versus log(1 – R^2) for study and fellow eyes at baseline and months 6, 12, 18, and 24, the resulting (best-fitting) pattern is linear (with slope = –1.22) and is in agreement with the theory regarding R^2 for regression and correlation models (Supplementary Fig. S3).[27] Further, as noted from the estimated residual errors (RMSE), the models seem to have similar performance at each time point (Table 3; Supplementary Table S1).

Although it is clear that the functional deficits in retinal disease are due to structural damage in the retina, determining the precise relationship between specific and measurable anatomical changes and vision is challenging.[3–6,8,9] As mentioned previously, the conventional approach to investigating this relationship is to pick one or more anatomic features and then to perform multivariable statistical and machine learning analyses to determine if any association with vision can be quantified.[11,12,14] As such, this approach is limited by the ability of the investigator to predetermine a potentially large set of retinal structures and features with the greatest likelihood of a meaningful relationship with visual function. In contrast, DL-based algorithms (and in particular CNNs) do not require any anatomic feature to be identified before the quantitative investigation. Instead, the DL algorithm evaluates the OCT image as a whole and learns directly from the images to identify features that enable the most accurate prediction of the outcome of interest. Compared with previously reported cross-validation results on a subset of 614 patients from HARBOR reporting $R^2 = 0.34$ when using known imaging features along with baseline BCVA to predict BCVA at month 12, our model achieved $R^2 = 0.45$ on the 924 patients in the tuning set and $R^2 = 0.40$ on the 126 patients in the internal validation test set.[14] In principle, this could lead to the identification of previously unappreciated anatomic features, or combinations of features, critical to visual function. However, one potential drawback of the DL approach is that the most predictive features of BCVA may be complex, nonlinear combinations of local and distal retinal structures. We performed saliency-based visualization, but these features did not lend themselves to easy interpretation. The identification of specific features remains an important area of future research. Here, our focus was to demonstrate the existence of a relationship between retinal OCT images and BCVA in patients with nAMD.

Separate DL models were able to predict BCVA value in the study eyes 12 months from the time of the baseline OCT measurement, with a moderate correlation of ~0.57 ($R^2 = 0.33$) (Table 5). It is interesting to note that, when added into a regression model that already contained baseline BCVA ($P < 0.001$), the OCT-based prediction (from baseline) remained highly statistically significant ($P < 0.001$). In this multivariable model, both predictors provide approximately equal information of future visual function, with a model $R^2 = 0.40$ (Table 5). If used as a stratification factor at baseline, this prediction model would translate into smaller/shorter trials at the same statistical power. Initially, we tried separate models for study and fellow eyes, but, surprisingly, the results were not significantly different in terms of predictive performance from the model trained on study and fellow eyes pooled together.

If verified and optimized for greater accuracy across both treated and untreated patients, this DL approach could have meaningful clinical utility. Measurement of BCVA is frequently cumbersome, requiring specialized resources for accurate refractive testing. Indeed, in the evaluation of retinal health outside of the clinic setting, the ability to augment visual function measurements with computer vision-based analysis of OCT images will likely prove valuable for screening and monitoring patients. For example, such an approach could aid remote consultations via telemedicine, where physicians could use DL data on the patient's current and future visual potential to support their clinical decisions. Furthermore, in clinical research, DL models that help predict future BCVA response could be used to support trial enrollment or trial stratification by focusing on individuals that are likely to benefit from treatment.

A limitation of this study was that the training of the algorithms was done on data from a single clinical trial, and it is unclear whether or not the results are generalizable to the overall population with nAMD. Additionally, our results may not be generalizable to macular changes secondary to other causes, such as myopic CNV, diabetic macular edema, or retinal vein occlusions. Our work is to be considered proof of feasibility, and future research will be needed to validate this DL model against data from other clinical trials, real-world data, and OCT images based on remote-monitoring technology. With further efforts, it could be possible to create more accurate and generalizable models capable of predicting visual function measures in the clinical setting.

This study demonstrated that DL algorithms could potentially help predict concurrent and future BCVA from OCT images in patients with nAMD. Further optimization of the presented models could help expand both our understanding of, and ability to effectively manage, this sight-threatening disease. In the future, the capability to quickly and accurately predict BCVA from OCT images could enable more efficient screening and detection of patients with early or progressive vision loss. Future research is needed to understand the utility of such DL algorithms in supporting clinicians and researchers in their clinical care and clinical research, respectively. Importantly, future research should also assess whether approaches founded on DL may provide insight into the biological bases that drive both structural and functional changes in nAMD.

Acknowledgments

The authors thank Anh Duong and Alex de Crespigny for project guidance; Elizabeth Li, Justin Lee, and Priya Ponnapalli for technical support and scientific discussions; and Mary K. Durbin, Archana Kolli, and Paolo Pochendorfer from Carl Zeiss Meditec, Inc. for technical support. Additional writing and editorial assistance was provided by Amy Lindsay, PhD, of Envision Pharma Group (funded by Genentech, Inc.).

Funding was provided by Genentech, Inc., a member of the Roche Group, for the study. Genentech, Inc. supported and contributed to all aspects of the study, including the study design, analyses, data interpretation, report writing, and decision to submit the manuscript for publication.

T.B., S.S.G., and J.R.W. designed the experiment; M.G.K., J.D., and S.S.G. conducted the experiment; J.D. and S.S.G. prepared the dataset; M.G.K. built the deep learning models; M.G.K., T.B., J.D., and S.S.G. analyzed and interpreted the data; M.G.K., T.B., S.S.G., and J.R.W. wrote the article; and M.G.K., T.B., J.D., J.J.H., S.S.G., and J.R.W. proofed and revised the article.

Disclosure: **M.G. Kawczynski,** Genentech, Inc. (E); **T. Bengtsson,** Genentech, Inc. (E); **J. Dai,** Genentech, Inc. (E); **J.J. Hopkins,** Genentech, Inc. (E); **S.S. Gao,** Genentech, Inc. (E); **J.R. Willis,** Genentech, Inc. (E)

References

1. Brown DM, Kaiser PK, Michels M, et al. Ranibizumab versus verteporfin for neovascular age-related macular degeneration. *N Engl J Med*. 2006;355:1432–1444.
2. Busbee BG, Ho AC, Brown DM, et al. Twelve-month efficacy and safety of 0.5 mg or 2.0 mg ranibizumab in patients with subfoveal neovascular age-related macular degeneration. *Ophthalmology*. 2013;120:1046–1056.
3. Comparison of Age-Related Macular Degeneration Treatments Trials (CATT) Research Group, Maguire MG, Martin DF, Ying GS, et al. Five-year outcomes with anti–vascular endothelial growth factor treatment of neovascular age-related macular degeneration: the Comparison of Age-Related Macular Degeneration Treatments Trials. *Ophthalmology*. 2016;123:1751–1761.
4. Bonini Filho MA, Witkin AJ. Outer retinal layers as predictors of vision loss. *Rev Ophthalmol*. Available at: https://www.reviewofophthalmology.com/article/outer-retinal-layers-as-predictors-of-vision-loss. Accessed August 6, 2019.
5. Charbel Issa P, Troeger E, Finger R, Holz FG, Wilke R, Scholl HP. Structure-function correlation of the human central retina. *PLoS One*. 2010;5:e12864.
6. Jaffe GJ, Martin DF, Toth CA, et al. Macular morphology and visual acuity in the Comparison of Age-Related Macular Degeneration Treatments Trials. *Ophthalmology*. 2013;120:1860–1870.
7. Schmidt-Erfurth U, Klimscha S, Waldstein SM, Bogunović H. A view of the current and future role of optical coherence tomography in the management of age-related macular degeneration. *Eye (Lond)*. 2017;31:26–44.
8. Schmidt-Erfurth U, Waldstein SM, Deak GG, Kundi M, Simader C. Pigment epithelial detachment followed by retinal cystoid degeneration leads to vision loss in treatment of neovascular age-related macular degeneration. *Ophthalmology*. 2015;122:822–832.
9. Sharma S, Toth CA, Daniel E, et al. Macular morphology and visual acuity in the second year of the Comparison of Age-Related Macular Degeneration Treatments Trials. *Ophthalmology*. 2016;123:865–875.
10. Aslam TM, Zaki HR, Mahmood S, et al. Use of a neural net to model the impact of optical coherence tomography abnormalities on vision in age-related macular degeneration. *Am J Ophthalmol*. 2018;185:94–100.
11. Mathew R, Richardson M, Sivaprasad S. Predictive value of spectral-domain optical coherence tomography features in assessment of visual prognosis in eyes with neovascular age-related macular degeneration treated with ranibizumab. *Am J Ophthalmol*. 2013;155:720–726.e1.
12. Muether PS, Hermann MM, Koch K, Fauser S. Delay between medical indication to anti-VEGF treatment in age-related macular degeneration can result in a loss of visual acuity. *Graefes Arch Clin Exp Ophthalmol*. 2011;249:633–637.
13. Rohm M, Tresp V, Müller M, et al. Predicting visual acuity by using machine learning in patients treated for neovascular age-related macular degeneration. *Ophthalmology*. 2018;125:1028–1036.
14. Schmidt-Erfurth U, Bogunovic H, Sadeghipour A, et al. Machine learning to analyze the prognostic value of current imaging biomarkers in neovascular age-related macular degeneration. *Ophthalmol Retina*. 2018;2:24–30.

15. Waldstein SM, Simader C, Staurenghi G, et al. Morphology and visual acuity in aflibercept and ranibizumab therapy for neovascular age-related macular degeneration in the VIEW trials. *Ophthalmology*. 2016;123:1521–1829.
16. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Available at: https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf. Accessed August 6, 2019.
17. Lee CS, Baughman DM, Lee AY. Deep learning is effective for classifying normal versus age-related macular degeneration OCT images. *Ophthalmol Retina*. 2017;1:322–327.
18. Rubin GS, West SK, Muñoz B, et al. A comprehensive assessment of visual impairment in a population of older Americans. The SEE Study. Salisbury Eye Evaluation Project. *Invest Ophthalmol Vis Sci*. 1997;38:557–568.
19. American Foundation for the Blind. Statistical snapshots from the American Foundation for the Blind. Available at: https://www.afb.org/research-and-initiatives/statistics. Accessed July 2, 2019.
20. Ho AC, Busbee BG, Regillo CD, et al. Twenty-four-month efficacy and safety of 0.5 mg or 2.0 mg ranibizumab in patients with subfoveal neovascular age-related macular degeneration. *Ophthalmology*. 2014;121:2181–2192.
21. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. New York, NY: Institute of Electrical and Electronics Engineers; 2016:770–778.
22. He K, Zhang X, Ren S, Sun J. Identity mappings in deep residual networks. In: Leibe B, Matas J, Sebe N, Welling M, eds. *Computer Vision–ECCV 2016. Lecture Notes in Computer Science*, Vol. 9908. Cham, Germany: Springer; 2016:630–645.
23. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdino R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res*. 2014;15:1929–1958.
24. Ng AY. Feature selection, L_1 vs. L_2 regularization, and rotational invariance. In: *Proceedings of the 21st International Conference on Machine Learning*. New York, NY: Association for Computing Machinery; 2004:78.
25. Liu L, Jiang H, He P, et al. On the variance of the adaptive learning rate and beyond. Available at: https://arxiv.org/pdf/1908.03265.pdf. Accessed August 4, 2020.
26. Patel PJ, Chen FK, Rubin GS, Tufail A. Intersession repeatability of visual acuity scores in age-related macular degeneration. *Invest Ophthalmol Vis Sci*. 2008;49:4347–4352.
27. Bland JM, Altman DG. Correlation in restricted ranges of data. *BMJ*. 2011;342:d556.