Deep-learning-based Prediction of Late Age-Related Macular Degeneration Progression

Qi Yan[1,*], Daniel E. Weeks[2,3], Hongyi Xin[1], Heng Huang[4], Anand Swaroop[5], Emily Y. Chew[6], Ying Ding[3,*], Wei Chen[1,2,3,*]

[1]Division of Pulmonary Medicine, Department of Pediatrics, Children's Hospital of Pittsburgh of UPMC, University of Pittsburgh, Pittsburgh, PA;

[2]Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, PA;

[3]Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA;

[4]Department of Electrical and Computer Engineering, Swanson School of Engineering, University of Pittsburgh, Pittsburgh, PA;

[5]Neurobiology Neurodegeneration and Repair Laboratory, National Eye Institute, National Institutes of Health, Bethesda, MD;

[6]Division of Epidemiology and Clinical Applications, National Eye Institute, National Institutes of Health, Bethesda, MD.

[*]These authors contributed equally

ABSTRACT

Age-related macular degeneration (AMD) is a leading cause of irreversible blindness worldwide. Previous genetic and epidemiological studies have shown that the AMD progression are affected by genetic variants and environmental factors. AMD severity is mainly diagnosed by color fundus images and recent studies have shown the success of machine learning methods in predicting AMD progression using image data. However, none of the studies have utilized both genetics and image data for predicting AMD progression. The Age-Related Eye Disease Study (AREDS), a large-scale clinical trial from the National Eye Institute, includes massive genome-wide genotyping data, longitudinal color fundus photographs, and disease severity assessment over a period of 12 years, providing an unprecedented opportunity to investigate prediction models for AMD progression. In this report, we jointly used genotypes and fundus images to dynamically predict an eye as having progressed to late AMD (e.g., whether the time to progression to late AMD for the eye exceeds 3 years from the current visit) with a modified deep convolutional neural network (CNN). In total, we used 31,262 color fundus images centered on the macula from 1,351 subjects with corresponding genotypes and phenotypes available at baseline and follow-up visits. The first part of this model was derived from a CNN to extract features. Our results showed that the color fundus photos coupled with genotypes could predict late AMD progression with an averaged area under the curve (AUC) value of 0.85 (95%CI: 0.83-0.86). The results using fundus images alone showed an AUC of 0.81 (95%CI: 0.80-0.83). We have validated our results in an independent dataset of 200 Caucasians extracted from UK Biobank and the results showed an AUC of 0.9 (95%CI: 0.85-0.94) for predicting whether the eye progresses to late AMD exceeding the 3 years. We implemented our model in a cloud-based application for individual risk assessment.

INTRODUCTION

Age-related macular degeneration (AMD) is the leading cause of blindness among older adults in Caucasians [1-3]. It is a progressive neurodegenerative disease influenced by both environmental and genetic risk factors [4-6]. AMD severity is mainly diagnosed by color fundus images in a clinical setting by ophthalmologists. Late AMD comes in two forms: (1) Geographic atrophy (GA) also known as dry AMD is characterized by a gradual degeneration and disappearance of retinal pigment epithelium, photoreceptor cells, and the choriocapillaris layer in the central retina; (2) choroidal neovascularization (CNV) also known as wet AMD is characterized by the growth of new, leaky blood vessels into the retina causing widespread photoreceptor loss and ultimately rapid decline in visual acuity [7]. Late AMD is distinguished from the early and intermediate stages by the growth of drusen size and/or pigmentary abnormalities at the macula region [8,9]. Some patients with early or intermediate stage AMD maintain their vision for a long time with slow disease progression over time, but others quickly progress to one or both forms of late AMD.

Genetics plays a critical role in AMD pathogenesis. Genome-wide association studies (GWAS) and sequencing studies have identified many variants that are associated with AMD [10-13]. For example, a total of 52 independent genetic markers including both common and rare variants from 34 loci were reported to have associations with AMD risk in a recent large-scale genome-wide association study (GWAS) by the International AMD Genomics Consortium [11]. In addition to the successes in identifying AMD-related genetic markers from case-control designs, a recent study of AMD progression risk using the Age-Related Eye Disease Study (AREDS) dataset [14] showed that some of the known AMD risk variants could also influence progression time to late AMD [12].

In parallel to genetic studies, machine learning methods, particularly deep convolutional neural networks (DCNN) have been useful in image recognition and classification in ophthalmology. CNN have been used for the aforementioned automated AMD grading, identifying diabetic retinopathy and cardiovascular risk factors from fundus images, and interpreting and segmenting optical coherence tomography (OCT) images [15-22]. As opposed to traditional machine learning approaches that rely on "feature engineering", which involves computing features explicitly defined by experts [23-25] , CNN can learn features directly from the images themselves. CNN is a family of deep learning techniques characterized by enabling the networks to contain many computation layers that can automatically, deeply and comprehensively learn features from lower-level structures to more generalized higher-level structures. Recently several studies have used the color fundus images to perform automated AMD grading [19-21] and estimation of future risk of AMD [26] by applying convolutional deep learning methods. However, none of these methods consider genetic data in the prediction model.

In addition to using fundus images for AMD grading, in conjunction with genotypes, fundus image data can be used to predict the probability of late AMD progression exceeding certain inquired durations. Since late AMD is irreversible, such prediction could urge potential patients to start preventative care beforehand and slow down the disease progression. The Age-Related Eye Disease Study (AREDS), a large-scale clinical trial from the National Eye Institute, includes massive genome-wide genotyping data, longitudinal color fundus photographs, and disease severity assessment over a period of 12 years, providing an unprecedented opportunity for us to investigate AMD progressing using both dynamic (fundus image) and static (genetics) information.

In this study, we jointly used genotypes and fundus images to predict an eye as having progressed to late AMD (which may never occur) within certain inquired durations from the current visit. The inquired duration was selected in advance, and it was relative to the time when the fundus image was taken, not to the time of the baseline visit. Specifically, for one eye, the inputs included one fundus image taken at the current visit and the genotypes of the corresponding subject, and the output was the probability that the time to late AMD exceeds the inquired duration. To our knowledge, this is the first time to jointly use genotypes and fundus images in a prediction model.

METHODS

Study Population and Phenotype Definition

The study subjects were from the AREDS study sponsored by the National Eye Institute [14], which was a long-term, multicenter, prospective study of AMD and age-related cataracts with a 12-year follow-up period to assess the risk factors and impact of daily supplements. Eligible subjects were between 55 and 80 years old at baseline and free of sight-threatening conditions other than cataract or AMD. In this study, we confined our analyses to the AMD data and only used Caucasian participants with genotype data and at least one follow-up visit.

In the analyses, we used 31,262 color fundus images centered above the macula from 1,351 subjects with corresponding genotypes and phenotypes available at baseline and follow-up visits. The detailed image taking procedure was described elsewhere [27]. The AREDS AMD scale [28], based upon severity score from 1 to 12, was adopted to determine whether the eye was in the late AMD stage or not, which was measured based on centralized grading of these fundus images obtained at each semi-annual/annual follow-up visit [28]. For each non-late AMD eye at each visit, we calculated its time period between the current visit and the time to late AMD, defined as the

time to the first visit when the severity score reached 9 (noncentral GA) or higher (10: central GA, 11: CNV, and 12: CNV and central GA). If the eye's severity score was less than 9 by the end of follow-up, the time to late AMD was treated as censored at the last visit.

Prediction of the Probability that the Time to Late AMD Progression Exceeds an Inquired Year

Since the exact times to late AMD of censored eyes were unknown and we only knew the eyes were not progressed at certain time points, in order to make the full use of the available information, we performed a set of binary predictions instead of predicting a continuous progression time. Specifically, we predicted the probability that the time to late AMD progression exceeds the inquired durations, 2, 3, 4, 5, 6 and 7 years from the current visit. This implies some eyes cannot be used for certain of these prediction models. For example, if one subject was censored at year 5.5 without having progressed to late AMD, this subject's eyes were used for inquired years 2, 3, 4 and 5, because it was known that the eyes had not progressed by those inquired years. However, these eyes could have progressed on either side of the inquired years of 6 and 7, which made this eye unusable for these two inquired years. In addition to the late AMD scores (i.e., 9-12), a severity score of 1 indicates little or no AMD-related changes, whereas scores 2 through 8 indicate early or intermediate AMD [28]. Other variables we considered were current visit age and baseline smoking status (never, former, current).

Replication dataset

In addition to AREDS, we extracted 100 Caucasian subjects labeled with AMD at baseline visit and 100 Caucasian subjects without AMD reported by the end of at least 3 years of follow-up from UK Biobank[29] as an independent test dataset. For samples labeled with AMD, we only kept the ones with clear drusen with the help from experienced image specialist. In other words, we

had 100 subjects who progressed to AMD within 3 years and 100 subjects who progressed to AMD after 3 years (it is possible that they never progressed to AMD). All the subjects needed to have genotypes and high-quality fundus images at baseline. Please note that not all of the AMD subjects had late AMD and some had early/intermediate stages of AMD. The outcome was not exactly the same as we used in the training process, although we could assume that the early/intermediate AMD subjects might progress to late AMD in a short time (e.g., within 3 years).

Genotype Data

DNA samples from consenting subjects in AREDS were collected and genotyped centrally by the International AMD Genomics Consortium, as described previously [11]. In brief, a custom-modified HumanCoreExome array by Illumina was used to obtain the genotypes followed by imputation with the 1000 Genomes Project reference panel (Phase I). In this study, we used 52 independent genetic variants from 34 loci that were either confirmed or newly discovered to have associations with AMD risk in a GWAS by the International AMD Genomics Consortium [11]. We used additive genotypes (i.e., 0: no minor allele; 1: one copy of the minor allele; 2: two copies of the minor allele).

Data Partitioning

The total of 31,262 original images were first randomly divided into a training set (90% of the subjects) and testing set (10% of the subjects). Then, the training set was further divided into 10 folds so that 10-fold cross-validation was performed with 9 folds for training and 1 fold for validation and this process was repeated 10 times. Because each subject includes multiple images over 12-year visits on both eyes, we performed this separation at the subject level, which means that images from the same subject were included in the same fold.

Data Augmentation

A data augmentation procedure was applied to increase the diversity of the training dataset, and thus to reduce the risk of overfitting the CNN. We applied several augmentations to each image before rescaling to a square. First, images were horizontally mirrored to mimic the left and right eye orientations of each image. Second, images were randomly cropped less than 10% on both height and width to mimic images with not well-centered macula. The augmented images were assigned the same labels as the corresponding original images. The purpose of image augmentation is to control for the overfitting problem by artificially creating training images, and these augmented images still preserve the key information of the image but are different from their original images. The proposed augmentation techniques were similar to the previous CNN study of fundus images [19]. After augmentation, all fundus images were resized to squares encompassing the macula and rescaled to 224×224 pixels.

Deep Learning Approach Combining Images and Genotypes

The first part of the model architecture is CNN used to extract features from fundus images (**Figure 1**). CNN is a special type of deep neural network that consists of many repeated processing layers that match the input image with successive convolutional filters to extract image features from low to high levels. A CNN as a member of deep neural networks that optimizes the weights of each layer using stochastic gradient descent via a backpropagation process. There can be millions of weights [30]. There are a number of existing CNN architectures designed for image processing [31]. In general, these architectures are similar. They are all comprised of sequentially convolutional and pooling layers. Each of the different architectures is best suited for specific problems. In this study, we used the Inception-v3 CNN architecture [32] to extract image features, which has been used for fundus images in several studies [19,21,22,33].

Additionally, we used pre-trained weights as the initial values to train our network, which were trained for general image classification using the ImageNet database [34] that contains thousands of different objects and millions of images. This scale of data is usually unavailable in medical image classification studies.

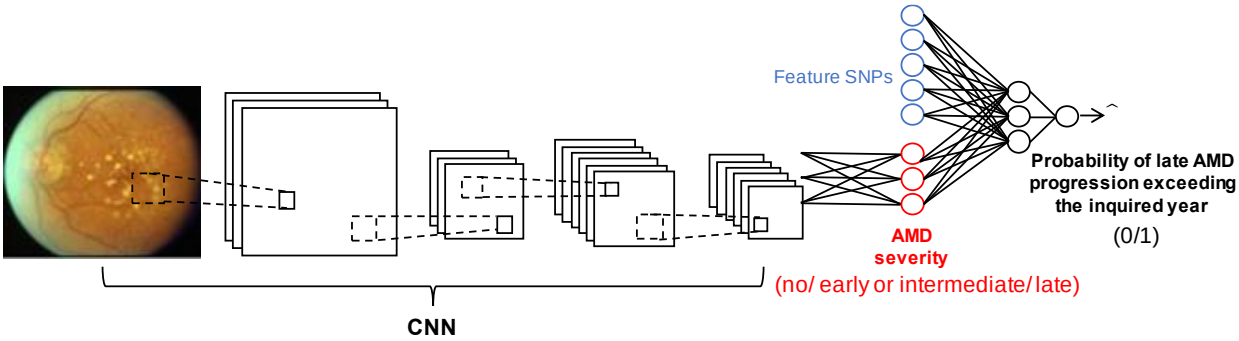

Figure 1. Convolutional neural network (CNN) of retinal fundus images along with feature SNPs and AMD severity for the prediction of late-AMD progression exceeding certain inquired years.

It was reported that the current AMD severity was the strongest predictor for the progression time to late AMD [35] and fundus images could be used to automatically grade AMD severity [19-21] with similar CNN architectures as presented here. Thus, after obtaining the output vector of the final convolutional layer, which contains all the information needed to understand the image, we fed these extracted image features to a fully connected layer to classify AMD severity (**Figure 1**). Furthermore, this severity viewed as the current AMD severity along with 52 independent genetic variants were fed to another fully connected layer to predict the time to late AMD development exceeding certain inquired years (**Figure 1**). During training, we used 3-level severity labels (e.g., no AMD when severity score is equal to 1; early or intermediate AMD when severity score is between 2 and 8; and late AMD when severity score is greater than or equal to 9). The network was built for any inquired year k, where k = 2, 3, 4, 5, 6, and 7.

The aforementioned model could be simplified to sub-models. In total, we considered four models: (1) using fundus images taken at the current visit to predict whether the eye's progression time to late AMD exceeds the inquired year; (2) using both fundus images and feature SNPs to predict whether the eye's progression time to late AMD exceeds the inquired year; (3) using fundus images to predict whether the eye's progression time to late AMD exceeds the inquired year as well as to classify the AMD severity at the current visit; (4) using both fundus images and feature SNPs to predict whether the eye's progression time to late AMD exceeds the inquired year as well as to classify the AMD severity at the current visit.

Our deep learning network was implemented by using Keras with TensorFlow [36]. During the training process, we first fixed the pre-trained weights in the CNN and updated the rest of the weights using an Adam optimizer with a learning rate of 0.001. The aforementioned 10-fold cross-validation was performed. Based on the average performance of the validation set, the best epoch was selected for testing set evaluation. For each fold, we trained the network for 20 epochs (loops of the entire training set) and selected the most suitable epoch for testing after 10-fold cross-validation. Furthermore, we set all weights as trainable and fine-tuned the network with a learning rate of 0.0001 and selected the best epoch out of 10 epochs also after 10-fold cross-validation. In this fine-tuning step, we selected a small total number of epochs to avoid updating the pre-trained CNN weights too much. All the training was conducted on a machine equipped with an EVGA GeForce RTX 2080 Ti 11Gb GPU and 128Gb available RAM.

Performance Metrics

For the stepwise binary predictions, we calculated the AUC (area under the curve) of Receiver Operator Characteristic (ROC) curves as the primary performance metric. Besides, Brier score [37] that is the squared error of a probabilistic prediction was used as another metric, and the lower

the Brier score the better the model predicts. The useful benchmark values for the Brier score are 33%, which corresponds to predicting the risk by a random number drawn from a Uniform distribution between 0 and 1. For the 3-level AMD severity classification, no AMD, early or intermediate AMD and late AMD were treated as levels 0, 1 and 2 (see supplementary text for details).

Visualizing Model Attention

To help understand what image features were learned and make the "black box" deep learning model more transparent, we generated saliency maps to highlight the regions that most contribute to the predicted values from the output layer for all trained models. The saliency is computed by the gradient of output value with respect to the input image. In other words, it could detect how a small change in the input image changes the output value. If these gradients have the same shape as the region of interest, it indicates that the attention of learning is on the right region. Thus, these gradients could be used to highlight input regions that result in the most change in the output prediction.

Data Availability

All the phenotype data and fundus images of AREDS participants required are available from dbGap (accession: phs000001.v3.p1). The genotype data on AREDS subjects has been reported earlier [11] and is available from dbGap (accession phs001039.v1.p1). The UK Biobank test dataset was obtained from UK Biobank (application number 43252).

RESULTS

Study Data Characteristics

Of the AREDS participants, 1,351 Caucasians who had at least one eye free of late AMD at baseline and at least one follow-up visit had all information on images and genotypes (**Table 1**). The baseline mean age was 68.8 (SD=5.0) years. About 56% (N=750) of participants were females. About 46% were never smokers (N=626), another 47% were former smokers (N=634) and 7% were current smokers (N=91). The participants had mean follow-up time of 10.3 (SD=1.6) years and they were followed up every 6 months in the first 6 years to every 1 year after year 6. 2,678 eyes of the 1,351 participants were not in the late AMD stage at baseline. These eyes had a low mean severity score at baseline of 3.9 (SD=3.2), because the majority of eyes had low baseline severity scores (54% eyes with baseline severity score 1-3, 23% eyes with 4-6 and 24% eyes with 7-8). Moreover, only 4% eyes with baseline severity 1-3 progressed to late AMD by the end of the follow-up time, 50% eyes with 4-6 progressed by the end of the follow-up time, and 92% eyes with 7-8 progressed by the end of the follow-up time. In addition, the number of useable fundus images (i.e., the fundus images of each eye at each visit with corresponding genotypes) for prediction decreased as the progression inquired year increased from 2 to 7 years due to the censored subjects (Table 1). The inquired years were defined as well as an illustrative example was provided in the Methods section. The number of useable fundus images for predicting whether the progression time to late AMD exceeded the inquired years were shown in Figure S1. Note that the useable fundus images included the ones at both baseline and the follow-up visits.

Table 1. Characteristics of the participants

	AREDS	Training	Test
Subject-level variables	1,351 subjects	1,223 subjects	128 subjects
Baseline age, year (mean ± SD)	68.8 ± 5.0	68.8 ± 5.0	68.5 ± 4.8
Female (N, %)	750 (55.5)	682 (55.8)	68 (53.1)
Follow-up time, (mean ± SD)	10.3 ± 1.6	10.2 ± 1.7	10.9 ± 1.0
Smoking (N, %)			
Never smoked	626 (46.3)	566 (46.3)	60 (46.9)
Former smoker	634 (46.9)	576 (47.1)	58 (45.3)
Current smoker	91 (6.7)	81 (6.6)	10 (7.8)
Eye-level variables	2,678 eyes	2,422 eyes	256 eyes
Baseline AMD severity score at eye-level			
Mean ± SD	3.9 ± 3.2	4.0 ± 3.2	3.9 ± 3.2
1-3 (n, %)	1,442 (53.8)	1,310 (54.1)	132 (51.6)
4-6 (n, %)	600 (22.5)	528 (21.8)	72 (28.1)
7-8 (n, %)	636 (23.7)	584 (24.1)	52 (20.3)
Progressed eyes with baseline severity			
1-3 (n, %)	50 (3.5)	48 (3.7)	2 (1.5)
4-6 (n, %)	300 (50.0)	260 (49.2)	40 (55.6)
7-8 (n, %)	585 (92.0)	537 (92.0)	48 (92.3)
Observation-level variables			
Fundus images used for prediction with progression cutoff			
2 years (n)	27,499	24,654	2,845
3 years (n)	25,862	23,170	2,692
4 years (n)	24,287	21,709	2,578
5 years (n)	22,435	20,041	2,394
6 years (n)	20,240	18,118	2,122
7 years (n)	18,066	16,172	1,894

Predicting Progression Time to Late AMD Exceeding Inquired Years Using Fundus Images Alone

First, we tested the ability of our proposed network to predict the progression time to late AMD exceeding inquired years using the AREDS fundus images alone (Figure S2A). The partitioning of training and testing datasets and the selection of late AMD progression inquired years were previously explained. The ROC and AUC (Figure 1) and Brier scores on the testing dataset are reported in Table 2. The performance of our CNN showed promising results even when using fundus images alone. In the testing dataset, the model achieved AUC range from 0.79 (95% CI

[confidence interval]: 0.78-0.81) to 0.84 (95% CI: 0.82-0.86) for progression inquired year 2 ~ 7 (Figure 2 and Table 2). Furthermore, in order to improve the interpretability of our model, we added a secondary output layer for current AMD severity between the final convolutional layer and the primary output for the progression time to late AMD (Figure S2B). The model achieved similar AUC range between 0.79 (95% CI: 0.77-0.81) and 0.84 (95% CI: 0.83-0.86) for progression inquired year 2 ~ 7 (Figure 2 and Table 2). Besides, the model automatically graded the AMD severity based on fundus images with an accuracy range of 0.56 to 0.60 for progression inquired year 2 ~ 7 (Table S2). Note that the random accuracy is 0.33. The density curves of predicted probability of having late AMD progression time before or after each of the six inquired years were generated (Figure S3 [the first and third columns]) to visually examine the accuracy of before and after inquired year prediction separately. The results showed that CNN could accurately predict the probability of having late AMD progression time exceeding the inquired year. However, although most of the eyes with progression time before the inquired year could be correctly predicted when Youden indices [38] (Table S3) were used as the thresholds to dichotomize the predictions, a sizeable number of eyes were falsely predicted as having a progression time exceeding the inquired year.

Table 2. AUC values (95% CI) and Brier scores (95% CI) of the prediction of probability of late-AMD progression exceeding the inquired years for four models

		2 years	3 years	4 years	5 years	6 years	7 years
AUC values	Image -> Risk*	0.81 (0.79-0.83)	0.81 (0.79-0.83)	0.81 (0.79-0.83)	0.79 (0.78-0.81)	0.83 (0.81-0.85)	0.84 (0.82-0.86)
	Image + Geno -> Risk	0.84 (0.83-0.86)	0.85 (0.83-0.86)	0.83 (0.82-0.85)	0.84 (0.82-0.85)	0.85 (0.83-0.86)	0.85 (0.83-0.87)
	Image -> AMDstate# -> Risk	0.81 (0.79-0.83)	0.80 (0.78-0.82)	0.79 (0.77-0.81)	0.82 (0.80-0.84)	0.82 (0.80-0.84)	0.84 (0.83-0.86)
	(Image -> AMDstate) + Geno -> Risk	0.85 (0.84-0.87)	0.86 (0.84-0.87)	0.86 (0.84-0.87)	0.85 (0.84-0.87)	0.85 (0.84-0.87)	0.85 (0.84-0.87)
Brier scores	Image -> Risk	0.13 (0.12-0.14)	0.14 (0.13-0.15)	0.15 (0.14-0.16)	0.16 (0.15-0.17)	0.15 (0.15-0.16)	0.16 (0.15-0.17)
	Image + Geno -> Risk	0.13 (0.12-0.15)	0.15 (0.14-0.16)	0.16 (0.15-0.17)	0.17 (0.16-0.18)	0.17 (0.16-0.18)	0.17 (0.16-0.18)
	Image -> AMDstate -> Risk	0.13 (0.12-0.14)	0.15 (0.14-0.16)	0.16 (0.15-0.17)	0.15 (0.14-0.16)	0.16 (0.16-0.17)	0.16 (0.15-0.17)
	(Image -> AMDstate) + Geno -> Risk	0.13 (0.12-0.14)	0.14 (0.13-0.15)	0.14 (0.14-0.16)	0.15 (0.14-0.16)	0.16 (0.15-0.17)	0.16 (0.15-0.17)

AUC 95% CI uses the DeLong method [39]; Brier score 95% CI uses bootstrap method.

* The probability of late-AMD progression exceeding the inquired years.

No, early or intermediate, or late AMD (3 levels).

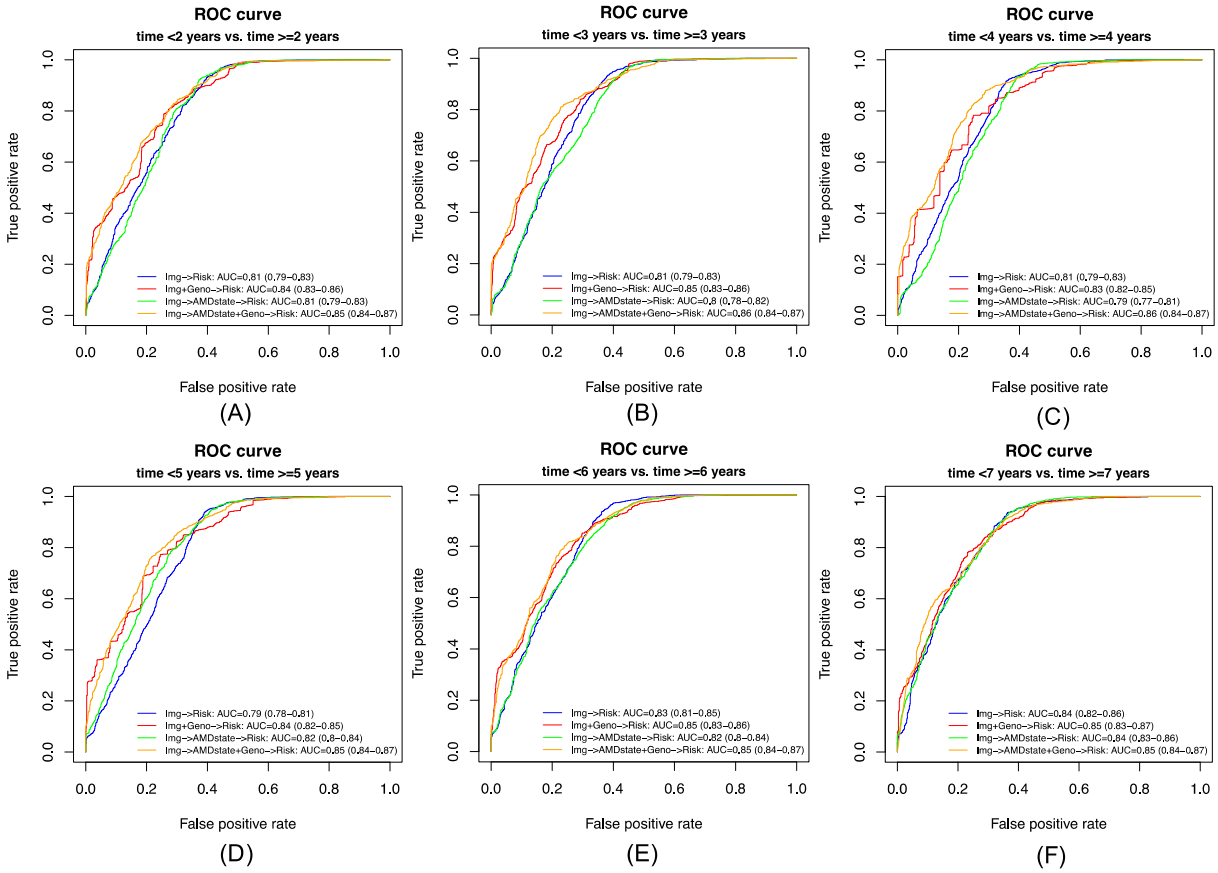

Figure 2. ROC curves of the prediction of late AMD progression time exceeding the inquired years for four models. The four models are (1) fundus images predicting late-AMD progression exceeding the inquired years; (2) fundus images + genotypes predicting late-AMD progression exceeding the inquired years; (3) fundus images both classifying current AMD severity and predicting late-AMD progression exceeding the inquired years; and (4) fundus images + genotypes both classifying current AMD severity and predicting late-AMD progression exceeding the inquired years. (A-F) inquired years from 2 to 7.

Model Interpretation for Supporting Clinical Decision

Next, we generated saliency maps to visualize the important regions that had the greatest impact on the model predictions using the fundus images only models. Generally, the CNN should be able to detect the macula region and make decisions based on the features (e.g., drusen) in this region. Three representative subjects of left eye fundus images with accompanying saliency maps for each inquired year prediction were shown in Figures 3-5. Subject #1 (Figure 3) had 4 visits and progressed to late AMD at year 4.8; Subject #2 (Figure 4) had 4 visits and did not

progress to late AMD by the end of 11.1-year follow-up; and Subject #3 (Figure 5) had 5 visits and was at late AMD status at baseline. Thus, Subject #1 who progressed in the middle of the follow-up was more challenging to predict than Subject #2 who did not progress and Subject #3 who progressed before baseline time. Figure 3 showed that Subject #1 had early/intermediate AMD at the first three visits and was labeled as late AMD at the fourth visit. Most models with different cut-off years gave accurate predictions. However, the model with inquired year equal to 3 tended to predict a long progression time when small drusen was observed. This was also the reason that the density curves with only images as predictors led to a second peak on the right side (blue curves in Figure S3). In other words, the samples that were falsely predicted to the category exceeding the inquired years were true labeled in the category before the inquired years. The results showed that genotypes could help to correct this misclassification. Figure 4 showed that Subject #2 had healthy macula at all visits and the models correctly predicted that the subject had a long progression time. Please note that the true labels were missing at visit year 5.8 for inquired years equal to 6 and 7, because this subject was censored and had 5.3 more follow-up years at visit year 5.8. Thus, at visit year 5.8, this subject could finally progress to late AMD before or after 6 more years, which was ambiguous to use this cutoff, same to the cutoff of year 7. Figure 5 showed that Subject #3 had late AMD since baseline, and all models could detect the drusen and correctly assign it to the right category.

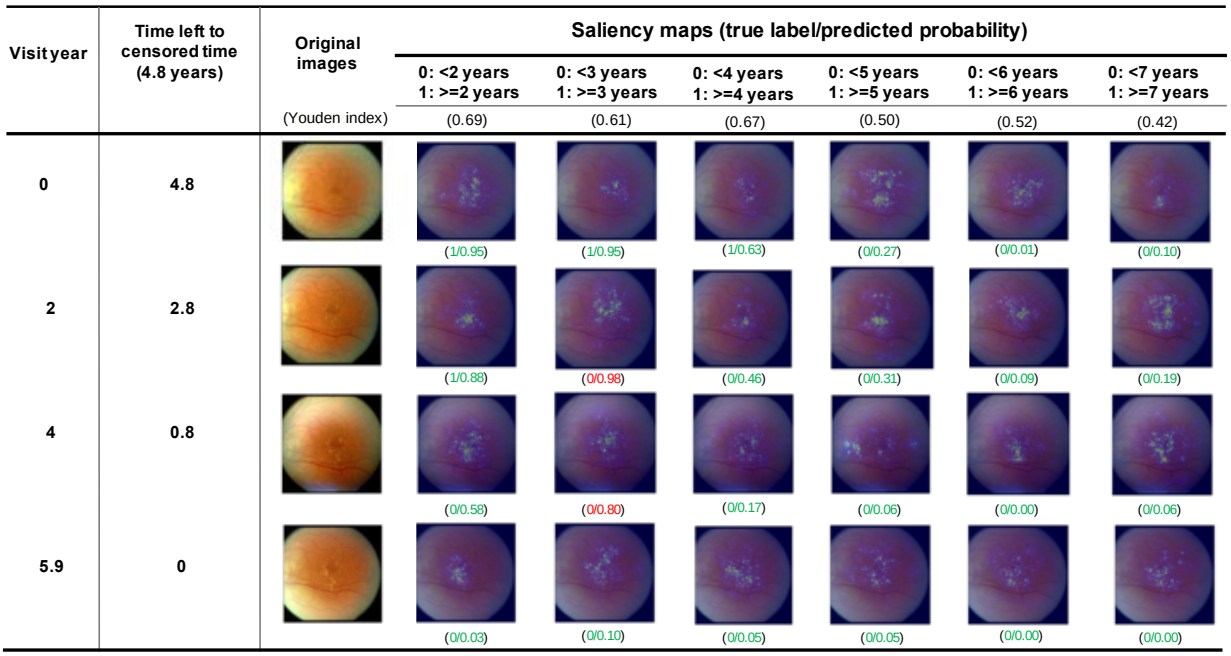

Figure 3. Saliency maps for left eye of Subject #1 over 5.9 years. This subject progressed to late AMD after 4.8 years of follow-up. The highlighted dots indicate the area that the CNN learned to make the decision. The first number in the parenthesis is the true label (1=not progressed, 0=progressed) and the second number is the estimated probability of late AMD progression time exceeding the inquired year relative to the current visit. The green numbers indicate accurate predictions and red numbers indicate inaccurate predictions using Youden indices (Table S3, Img -> Risk) as the thresholds to dichotomize the predictions.

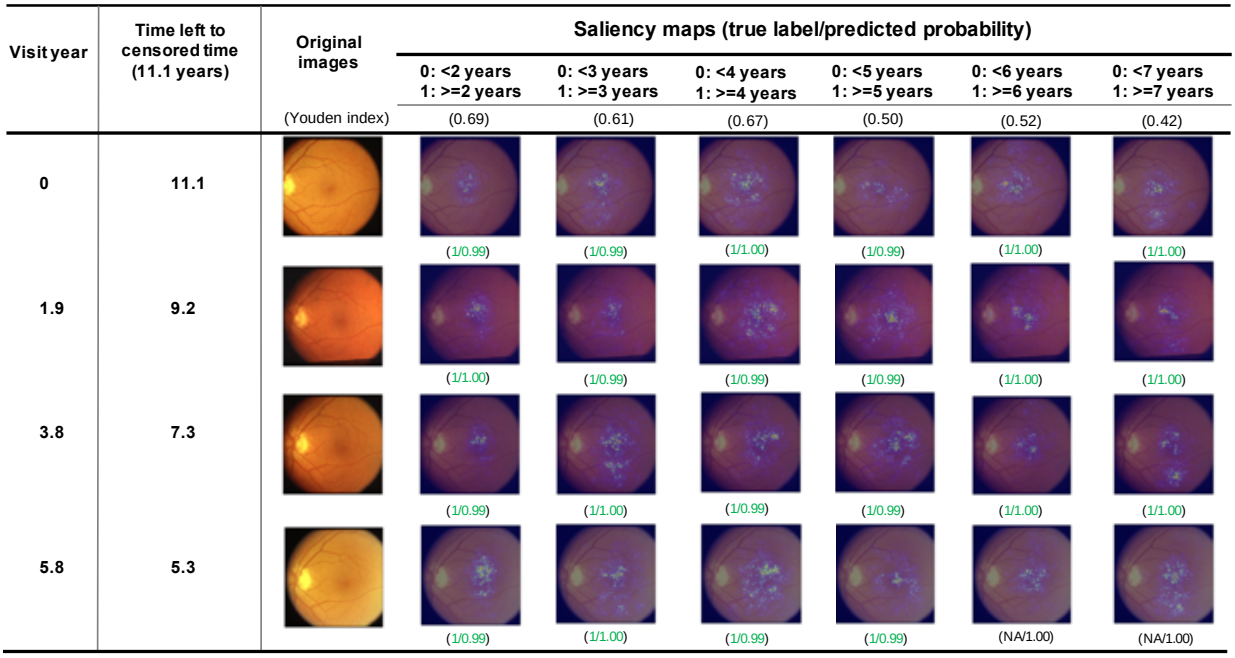

Figure 4. Saliency maps for left eye of Subject #2 over the first 5.8 years. This subject was censored after 11.1 years of follow-up. The highlighted dots indicate the area that the CNN learned to make the decision. The first number in the parenthesis is the true label (1=not progressed, 0=progressed, NA=progression status unknown) and the second number is the estimated probability of late AMD progression time exceeding the inquired year relative to the current visit. The green numbers indicate accurate predictions and red numbers indicate inaccurate predictions using Youden indices (Table S3, Img -> Risk) as the thresholds to dichotomize the predictions.

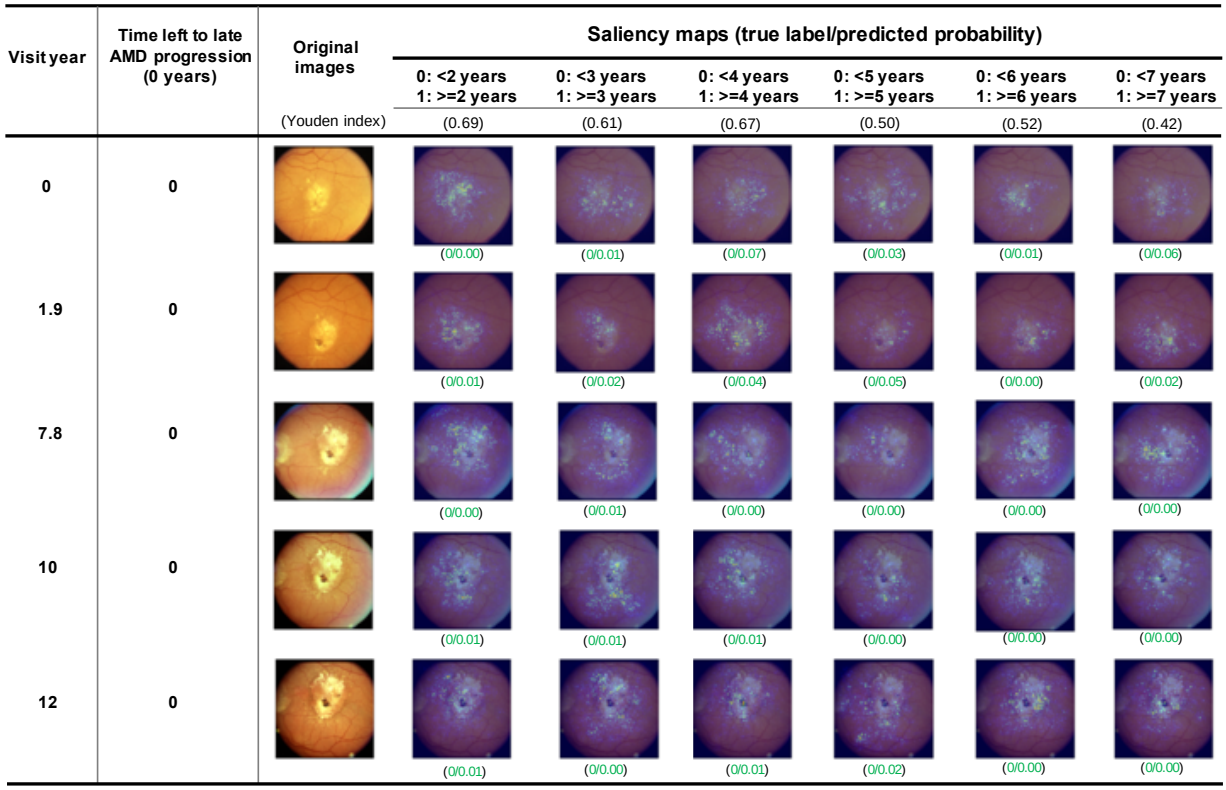

Figure 5. Saliency maps for left eye of Subject #3 over 12 years. This subject developed late AMD before enrollment. The highlighted dots indicate the area that the CNN learned to make the decision. The first number in the parenthesis is the true label (1=not progressed, 0=progressed) and the second number is the estimated probability of late AMD progression time exceeding the inquired year relative to the current visit. The green numbers indicate accurate predictions and red numbers indicate inaccurate predictions using Youden indices (Table S3, Img -> Risk) as the thresholds to dichotomize the predictions.

Predicting Progression Time to Late AMD Exceeding Inquired Years Using Fundus Images along with genotypes.

In addition to fundus images, we added 52 AMD associated independent genetic variants reported by the International AMD Genomics Consortium [11] to the model (Figure S2C). Similarly, we tested the scenarios with and without AMD severity as a secondary output (Figure S2D). When evaluating the performance using the test dataset, the models with AMD severity as a secondary output achieved slightly higher AUC (range between 0.85 [95% CI: 0.84-0.87] and 0.86 [95% CI: 0.84-0.87] for progression inquired year 2 ~ 7) than without AMD severity (range between 0.83 [95% CI: 0.82-0.85] and 0.85 [95% CI: 0.83-0.86], Figure 2 and Table 2). The

accuracy for AMD severity grading was from 0.57 to 0.60 (Table S2). The AMD severity as a secondary output helped to better explain the prediction model that the fundus images could be used for automated AMD severity grading, and then the AMD severity at the current visit could predict late AMD progression exceeding the inquired years. Without AMD severity, we only knew fundus images could predict late AMD progression exceeding the inquired years, but missed the related features (i.e., current AMD severity) in between. These results also indicated that genetics could noticeably improve the AUCs for AMD progression time (Table 2). The density curves of predicted probability of having AMD progression time before or after the six inquired years (Figure S3) showed that the addition of genetics in the neural network could correct the falsely predicted eyes from the group of after the inquired year to the correct group of before the inquired year. This is evidenced by that the blue curves on the right (Figure S3) were much lower in models with genetics than models without genetics.

Replication Study Using Independent UK Biobank Dataset

In addition to the test dataset we generated from dbGaP, we extracted a set of 200 Caucasian subjects from the UK Biobank[29] as an independent test dataset, and used 3 years as the inquired year. The model (4) showed an AUC of 0.9 (95%CI: 0.85-0.94) for predicting whether the eye progresses to late AMD exceeding the 3 years (Figure S4). Again, please note that we treated the AMD patients from UK Biobank as late AMD patients, although some of them could be at early/intermediate stages.

Implementation and Availability

The prediction models are available at https://github.com/QiYanPitt/AMDprogressCNN and a web-based application is also available at http://www.pitt.edu/~qiy17/amdprediction.html.

DISCUSSION

Our results show that the application of CNN to retinal fundus images coupled with genotypes can be used to predict the probability of late AMD progression exceeding the inquired years and diagnose the current AMD severity. The results indicate that the fundus images alone can predict late AMD progression exceeding certain inquired years with reasonable accuracy. The addition of secondary inputs of genotypes can noticeably improve accuracy (Table 2). The addition of a secondary output of AMD severity further improve the model interpretability. We also conducted an additional model using current age and baseline smoking status as predictors in addition to fundus images and genetics (results not shown), but they did not improve the prediction performance, probably because the age and smoking effects were already reflected in fundus images. Although the training of a deep learning model is computationally intensive, once the model is trained and weights are saved, the prediction of a new subject takes only a few seconds.

The goal of our study is different from previous ones, which mainly used fundus images for automated grading of AMD severity scores. Since the severity score is graded by ophthalmologists merely based on the fundus images, those studies did not engage other phenotypes or genotypes, although it is known that AMD is associated with age, smoking status and a number of genetic variants. We believe that the prediction of late AMD progression time is more useful as compared to the current severity score for patients to start preventative care early. Furthermore, the automated classification using directly measured variables would also help reduce the discrepancy among human graders and reduce costs of large-scale image assessment projects.

We used the pre-designed CNN architecture and pre-trained weights as our initial values, which are used for general image classification of thousands of objects based on millions of images. For our specific fundus image task, this helps the model learn more accurately with less data, because the existing CNN architecture and trained weights can identify simple features (e.g., edges and orientation) and further combine them to more complex objects, which mimics the way our visual system works. Analogously, our proposed model could mimic the way that ophthalmologists interpret the fundus images by identifying macula, drusen, pigmentary changes, etc. The saliency maps indicate that our model pays attention to assess features in the macula region in the retina to predict late AMD progression exceeding certain inquired years, which is as expected. Furthermore, the model might capture some unknown features that are important for AMD progression, but neglected by ophthalmologists.

Our study has some limitations. First, this study mainly relies on the AREDS dataset. Although an independent dataset from UK Biobank is used, the available outcome is any AMD rather than late AMD. Even though there are no other large longitudinal AMD studies with both fundus images and genotypes available, it would be beneficial to evaluate our models on a separately collected dataset. Second, AREDS contains a large number of normal to early AMD eyes at the beginning of the study and many of them did not progress to the late AMD by the end of the follow-up. Although we used stepwise binary predictions instead of predicting a continuous time to try to fully use the censored eyes, we still lose some of them. Another potential limitation is that only fundus images were used in this study. It would be desirable to have a coherent prediction by using multiple types of images (e.g., optical coherence tomography and fundus autofluorescence images). In addition, the fundus images in AREDS were collected on both eyes over several years from the same participants. In the current study, the covariance between two eyes or between different visits was not considered. Although it is hard to consider the

information of correlated images using deep learning approaches, incorporating such covariance in the model could increase the prediction accuracy, which might be implemented by modifying the loss function at the output layer from the neural network.

In summary, this study showed that deep learning approaches could be used to automatically predict late AMD progression exceeding certain inquired years and classify the current AMD severity stages. The joint use of fundus images and genotypes can achieve good prediction accuracy. The deep learning methods could serve an important role in decision support systems for eye services by reducing assessment time, workload and financial burden by automated analysis. Such automated analysis identifying individuals who should be referred to a specialist could become increasingly acceptable to both patients and ophthalmologists. We also developed a web-based application at http://www.pitt.edu/~qiy17/amdprediction.html. To our knowledge, this is the first cloud-based prediction website for AMD with deep learning techniques. In addition to distinguishing retinal pathologies using the fundus images, such as AMD, this study can be extended to be applicable to other diseases associated with images, genotypes and phenotypes (e.g., Alzheimer's disease).

ACKNOWLEDGEMENTS

The independent validation has been conducted using the UK Biobank Resource under Application Number 43252.

References

1. Stark, K. *et al.* The German AugUR study: study protocol of a prospective study to investigate chronic diseases in the elderly. *BMC Geriatr* **15**, 130 (2015).
2. Congdon, N. *et al.* Causes and prevalence of visual impairment among adults in the United States. *Arch Ophthalmol* **122**, 477-85 (2004).
3. Swaroop, A., Chew, E.Y., Rickman, C.B. & Abecasis, G.R. Unraveling a multifactorial late-onset disease: from genetic susceptibility to disease mechanisms for age-related macular degeneration. *Annu Rev Genomics Hum Genet* **10**, 19-43 (2009).
4. Fritsche, L.G. *et al.* Age-related macular degeneration: genetics and biology coming together. *Annu Rev Genomics Hum Genet* **15**, 151-71 (2014).
5. Ratnapriya, R. & Chew, E.Y. Age-related macular degeneration-clinical review and genetics update. *Clin Genet* **84**, 160-6 (2013).
6. Grassmann, F., Fauser, S. & Weber, B.H. The genetics of age-related macular degeneration (AMD)--Novel targets for designing treatment options? *Eur J Pharm Biopharm* **95**, 194-202 (2015).
7. Holz, F.G. *et al.* Progression of geographic atrophy and impact of fundus autofluorescence patterns in age-related macular degeneration. *Am J Ophthalmol* **143**, 463-72 (2007).
8. Swaroop, A., Branham, K.E., Chen, W. & Abecasis, G. Genetic susceptibility to age-related macular degeneration: a paradigm for dissecting complex disease traits. *Hum Mol Genet* **16 Spec No. 2**, R174-82 (2007).
9. Ferris, F.L., 3rd *et al.* Clinical classification of age-related macular degeneration. *Ophthalmology* **120**, 844-51 (2013).
10. Fritsche, L.G. *et al.* Seven new loci associated with age-related macular degeneration. *Nat Genet* **45**, 433-9, 439e1-2 (2013).
11. Fritsche, L.G. *et al.* A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nat Genet* **48**, 134-43 (2016).
12. Yan, Q. *et al.* Genome-wide analysis of disease progression in age-related macular degeneration. *Hum Mol Genet* **27**, 929-940 (2018).
13. Chen, W. *et al.* Genetic variants near TIMP3 and high-density lipoprotein-associated loci influence susceptibility to age-related macular degeneration. *Proc Natl Acad Sci U S A* **107**, 7401-6 (2010).
14. Age-Related Eye Disease Study Research, G. The Age-Related Eye Disease Study (AREDS): design implications. AREDS report no. 1. *Control Clin Trials* **20**, 573-600 (1999).
15. Abramoff, M.D. *et al.* Improved Automated Detection of Diabetic Retinopathy on a Publicly Available Dataset Through Integration of Deep Learning. *Invest Ophthalmol Vis Sci* **57**, 5200-5206 (2016).
16. Gulshan, V. *et al.* Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **316**, 2402-2410 (2016).
17. Kermany, D.S. *et al.* Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell* **172**, 1122-1131 e9 (2018).
18. Devalla, S.K. *et al.* A Deep Learning Approach to Digitally Stain Optical Coherence Tomography Images of the Optic Nerve Head. *Invest Ophthalmol Vis Sci* **59**, 63-74 (2018).

19. Grassmann, F. *et al.* A Deep Learning Algorithm for Prediction of Age-Related Eye Disease Study Severity Scale for Age-Related Macular Degeneration from Color Fundus Photography. *Ophthalmology* **125**, 1410-1420 (2018).
20. Burlina, P.M. *et al.* Automated Grading of Age-Related Macular Degeneration From Color Fundus Images Using Deep Convolutional Neural Networks. *JAMA Ophthalmol* **135**, 1170-1176 (2017).
21. Peng, Y. *et al.* DeepSeeNet: A deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs. *Ophthalmology* (2018).
22. Poplin, R. *et al.* Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering* **2**, 158-164 (2018).
23. Abramoff, M.D., Garvin, M.K. & Sonka, M. Retinal imaging and image analysis. *IEEE Rev Biomed Eng* **3**, 169-208 (2010).
24. Yousefi, S. *et al.* Learning from data: recognizing glaucomatous defect patterns and detecting progression from visual field measurements. *IEEE Trans Biomed Eng* **61**, 2112-24 (2014).
25. Chaudhuri, S., Chatterjee, S., Katz, N., Nelson, M. & Goldbaum, M. Detection of blood vessels in retinal images using two-dimensional matched filters. *IEEE Trans Med Imaging* **8**, 263-9 (1989).
26. Burlina, P.M. *et al.* Use of Deep Learning for Detailed Severity Characterization and Estimation of 5-Year Risk Among Patients With Age-Related Macular Degeneration. *JAMA Ophthalmol* **136**, 1359-1366 (2018).
27. Age-Related Eye Disease Study Research, G. The Age-Related Eye Disease Study system for classifying age-related macular degeneration from stereoscopic color fundus photographs: the Age-Related Eye Disease Study Report Number 6. *Am J Ophthalmol* **132**, 668-81 (2001).
28. Davis, M.D. *et al.* The Age-Related Eye Disease Study severity scale for age-related macular degeneration: AREDS Report No. 17. *Arch Ophthalmol* **123**, 1484-98 (2005).
29. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12**, e1001779 (2015).
30. Angermueller, C., Parnamaa, T., Parts, L. & Stegle, O. Deep learning for computational biology. *Mol Syst Biol* **12**, 878 (2016).
31. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436-44 (2015).
32. Szegedy, C., Vanhoucke, V., Ioffe, S. & Shlens, J. Rethinking the inception architecture for computer vision. *CVPR* **preprint arXiv:1512.00567**(2016).
33. Christopher, M. *et al.* Performance of Deep Learning Architectures and Transfer Learning for Detecting Glaucomatous Optic Neuropathy in Fundus Photographs. *Sci Rep* **8**, 16685 (2018).
34. Russakovsky, O. *et al.* ImageNet Large Scale Visual Recognition Challenge. *IJCV* **preprint arXiv:1409.0575**(2015).
35. Ding, Y. *et al.* Bivariate Analysis of Age-Related Macular Degeneration Progression Using Genetic Risk Scores. *Genetics* **206**, 119-133 (2017).
36. Chollet, F. Keras. *GitHub repository*, *https://github.com/keras-team/keras* (2015).
37. Graf, E., Schmoor, C., Sauerbrei, W. & Schumacher, M. Assessment and comparison of prognostic classification schemes for survival data. *Stat Med* **18**, 2529-45 (1999).
38. Youden, W.J. Index for rating diagnostic tests. *Cancer* **3**, 32-5 (1950).
39. DeLong, E.R., DeLong, D.M. & Clarke-Pearson, D.L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837-45 (1988).