Deep Learning for Predicting Refractive Error From Retinal Fundus Images

Avinash V. Varadarajan,[1] Ryan Poplin,[1] Katy Blumer,[1] Christof Angermueller,[1] Joe Ledsam,[2] Reena Chopra,[3] Pearse A. Keane,[3] Greg S. Corrado,[1] Lily Peng,[1] and Dale R. Webster[1]

[1]Google Research, Google, Inc., Mountain View, California, United States
[2]Google DeepMind, Google, Inc., London, United Kingdom
[3]NIHR Biomedical Research Centre for Ophthalmology, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom

Correspondence: Lily Peng, Google Research, 1600 Amphitheatre Way, Mountain View, CA 94043, USA; lhpeng@google.com.

AVV and RP contributed equally to the work presented here and should therefore be regarded as equivalent authors.

LP and DRW contributed equally to the work presented here and should therefore be regarded as equivalent authors.

Submitted: January 18, 2018
Accepted: April 26, 2018

Citation: Varadarajan AV, Poplin R, Blumer K, et al. Deep learning for predicting refractive error from retinal fundus images. *Invest Ophthalmol Vis Sci*. 2018;59:2861–2868. https://doi.org/10.1167/iovs.18-23887

Purpose. We evaluate how deep learning can be applied to extract novel information such as refractive error from retinal fundus imaging.

Methods. Retinal fundus images used in this study were 45- and 30-degree field of view images from the UK Biobank and Age-Related Eye Disease Study (AREDS) clinical trials, respectively. Refractive error was measured by autorefraction in UK Biobank and subjective refraction in AREDS. We trained a deep learning algorithm to predict refractive error from a total of 226,870 images and validated it on 24,007 UK Biobank and 15,750 AREDS images. Our model used the "attention" method to identify features that are correlated with refractive error.

Results. The resulting algorithm had a mean absolute error (MAE) of 0.56 diopters (95% confidence interval [CI]: 0.55–0.56) for estimating spherical equivalent on the UK Biobank data set and 0.91 diopters (95% CI: 0.89–0.93) for the AREDS data set. The baseline expected MAE (obtained by simply predicting the mean of this population) was 1.81 diopters (95% CI: 1.79–1.84) for UK Biobank and 1.63 (95% CI: 1.60–1.67) for AREDS. Attention maps suggested that the foveal region was one of the most important areas used by the algorithm to make this prediction, though other regions also contribute to the prediction.

Conclusions. To our knowledge, the ability to estimate refractive error with high accuracy from retinal fundus photos has not been previously known and demonstrates that deep learning can be applied to make novel predictions from medical images.

Keywords: deep learning, machine learning, refractive error, retinal imaging

Uncorrected refractive error is one of the most common causes of visual impairment worldwide.[1] The refractive error of the eye is determined by several factors. Axial ametropia, which is ametropia related to ocular length, is considered to be the main source of spherical refractive error. An eye with an optical system too powerful for its axial length is regarded as "myopic," whereas the eye that is too weak is known as "hypermetropic." The crystalline lens and curvature of the cornea balance the optical system of the eye and also contribute to both spherical refractive error and astigmatic properties of the eye.[2]

The prevalence of refractive error is increasing, particularly myopic errors in Western and Asian populations.[3] Although largely treatable with prescription spectacles or contact lenses, the vast majority of those affected by refractive error live in low-income countries with minimal access to eye care and therefore may not receive even this noninvasive treatment.[4]

Novel and portable instruments, such as smartphone attachments to image the fundus[5] or apps to measure visual acuity,[6] offer a low-cost method of screening and diagnosing eye disease in the developing world. They have shown promise in the assessment of diabetic retinopathy[7] and the optic nerve[8] but are limited by their requirement for expert graders to interpret the images.

Artificial intelligence (AI) has shown promising results in the diagnosis and interpretation of medical imaging. In particular a form of AI known as deep learning allows systems to learn predictive features directly from the images from a large data set of labeled examples without specifying rules or features explicitly.[9] Recent applications of deep learning to medical imaging have produced systems with performance rivaling medical experts for detecting a variety of diseases, including melanoma,[10] diabetic retinopathy,[11,12] and breast cancer lymph node metastases.[13,14] Deep learning can also characterize signals that medical experts cannot typically extract from images alone, such as age, gender, blood pressure, and other cardiovascular health factors.[15] Despite the high accuracy of the algorithms produced by deep learning, the number of free parameters of the system makes it difficult to understand exactly which predictive features have been learned by the system. The ability to infer human interpretable features from a trained convolutional neural network is critical not only for building trust in the system, but it also enables more targeted hypothesis generation for understanding the underlying mechanism of disease. In fact, much of the previous work in ophthalmic research and clinical practice has relied heavily on a sophisticated process of guess and test: first generate rules or hypotheses of what features are most predictive of a desired outcome and then test these assumptions. With deep learning,

Table 1. Population Characteristics of Subjects in the UK Biobank and AREDS Data Sets

Characteristics	Development Set		Clinical Validation Set	
	UK Biobank	AREDS	UK Biobank	AREDS
Number of subjects	48,101	4,128	12,026	500
Number of images	96,081	130,789	24,007	15,750
Mean age at imaging visit(s), y (SD)	56.8 (8.2)	73.8 (4.92)	56.9 (8.2)	73.83 (5.22)
Sex, % male	44.9	44.3	44.9	42.8
Ethnicity	Black, 1.2%	Black, 3.7%	Black, 1.3%	Black, 4.0%
	Asian/PI, 3.4%	Asian/PI, 0.2%	Asian/PI, 3.6%	Asian/PI, 0.2%
	White, 90.6%	White, 95.7%	White, 90.1%	White, 95.2%
	–	Hispanic, 0.3%	–	Hispanic, 0.4%
	Other, 4.1%	Other, 0.2%	Other, 4.2%	Other, 0.2%
	Unknown, 0.7%	–	Unknown, 0.8%	–
Mean SE, diopters (SD)	−0.38 (2.63)	0.67 (2.00)	−0.34 (2.57)	0.60 (2.08)
Severe myopia, SE worse than −6.00 D	3.9%	0.7%	3.8%	0.6%
Moderate myopia, SE −3.00 D to −6.00 D	9.6%	4.0%	9.2%	5.4%
Mild myopia, SE up to −3.00 D	33.7%	24.1%	33.5%	23.8%
Mild hypermetropia, SE up to +2.00 D	41.1%	50.1%	41.7%	47.0%
Moderate hypermetropia, SE +2.00 to +5.00 D	9.6%	19.9%	9.8%	21.8%
Severe hypermetropia, SE worse than +5.00 D	1.3%	1.1%	1.2%	1.4%
Unknown SE	0.8%	0.0%	0.8%	0.0%

The SE values shown are the averaged value over both eyes in the case of the UK Biobank data set and the averaged value over both eyes across all the visits in the AREDS data set. PI, Pacific Islander.

one can first ask the network to predict the outcome of interest and then apply attention techniques to identify the regions of the image that is most predictive for the outcome of interest.

In this study, a deep learning model[16] was trained to predict the refractive error from fundus images using two different data sets. Refractive error, particularly axial ametropia, is associated with characteristic changes in the fundus, including the relative geometry and size of features at the retina due to the curvature of the eye. This has been well studied, particularly in myopic eyes that have a longer axial length.[17] Attention techniques were used to visualize and identify new image features associated with the ability to make predictions. This method may help us interpret the model to understand which retinal landmarks may contribute to the etiology of ametropia.

Methods

Data Sets

We used two data sets in this study: UK Biobank and Age-Related Eye Disease Study (AREDS). UK Biobank is an ongoing observational study that recruited 500,000 participants between 40 and 69 years old across the United Kingdom between 2006 and 2010. Each participant completed lifestyle questionnaires, underwent a series of health measurements, provided biological samples,[18] and were followed up for health outcomes. Approximately 70,000 participants underwent ophthalmologic examination, which included an assessment of refractive error using an autorefraction device (RC5000; Tomey Corp., Nagoya, Japan) as well as paired nonmydriatic optical coherence tomography (OCT) and 45-degree retinal fundus imaging using a three-dimensional OCT device (OCT-1000 Mark 2; Topcon Corp., Tokyo, Japan). Participants who had undergone any eye surgery, including bilateral cataract surgery, were excluded from participating in the ophthalmologic exams because this meant their primary refractive error status could not be determined.

The AREDS was a clinical trial in the United States that investigated the natural history and risk factors of age-related macular degeneration and cataracts. The trial enrolled participants between 1992 and 1998 and continued clinical follow-up until 2001 at 11 retinal specialty clinics. The study was approved by an independent data and safety monitoring committee and by the institutional review board for each clinical center. A total of 4757 participants aged 55 to 80 years at enrollment were followed for a median of 6.5 years.[19] As a part of an ophthalmologic exam, the participants underwent subjective refraction as well as color fundus photography at baseline and at subsequent visits. Briefly, the protocol for refraction involved retinoscopy and then further refinement with subjective refraction. Thirty-degree–field color fundus photographs were acquired with a fundus camera (Zeiss FF-series; Carl Zeiss, Oberkochen, Germany) using a reading center–approved transparency film.[20] For each visit in which refraction was performed, the corresponding macula-centered photos were used in this study.

A summary metric for refractive error, known as the spherical equivalent (SE), can be calculated using the formula spherical power + 0.5*cylindrical power. SE was available for both the UK Biobank and AREDS data set, but spherical power and cylindrical power were only available in the UK Biobank data set.

Each data set was split into a development set and a clinical validation set, which was not accessed during model development (Table 1). The division of development and clinical validation sets was done by subject.

Development of the Algorithm

A deep neural network model is a sequence of mathematical operations, often with millions of parameters (weights),[21] applied to input, such as pixel values in an image. Deep learning is the process of learning the right parameter values ("training") such that this function performs a given task, such as generating a prediction from the pixel values in a retinal fundus photograph. TensorFlow,[22] an open-source software library for deep learning, was used in the training and evaluation of the models.

The development data set was divided into two parts: a "train" set and a "tune" set. The tune set is also commonly

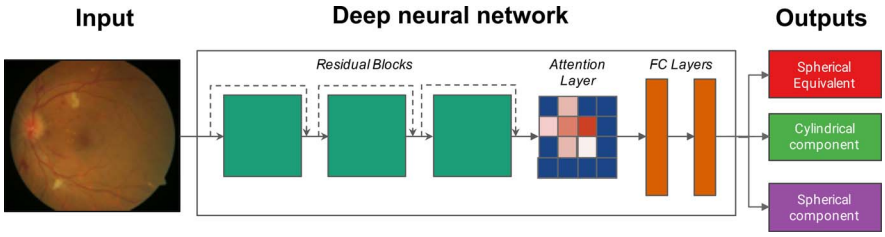

FIGURE 1. Overview diagram. Fundus images form the input of a deep neural network consisting of three residual blocks, an attention layer to learn the most predictive eye features, and two fully connected layers. Model outputs are SE, cylindrical component, and spherical component. Model parameters are learned in a data-driven manner by showing input-output examples.

called the "validation" set, but to avoid confusion with a clinical validation set (which consists of data on which the model did not train), we called it the tune set. The train, tune, and clinical validation data sets were divided by subject. During the training process, the parameters of the neural network were initially set to random values. Then for each image, the prediction given by the model was compared to the known label from the training set and parameters of the model were then modified slightly to decrease the error on that image. This process, known as stochastic gradient descent, was repeated for every image in the training set until the model "learned" how to accurately compute the label from the pixel intensities of the image for all images in the training set. The tuning data set was a random subset of the development data set that was not used to train the model parameters, but rather was used as a small evaluation data set for tuning the model. This tuning set comprised 10% of the UK Biobank data set and 11% of the AREDS data set. With appropriate tuning and sufficient data, the resulting model was able to predict the labels (e.g., refractive error) on new images. In this study, we designed a deep neural network that combines a ResNet[16] and a soft-attention[23] architecture (Fig. 1). Briefly, the network consists of layers to reduce the size of the input image, three residual blocks[16] to learn predictive image features, a soft-attention layer[23] to select the most informative features, and two fully connected layers to learn interactions between the selected features.

Prior to training, we applied an image-quality filter algorithm to exclude images of poor quality, which excluded approximate 12% of the UK Biobank data set. Because the vast majority of the AREDS images were of good quality, we did not exclude any of the AREDS images. The image-quality algorithm was a convolutional neural net trained on 300 manually labeled images to predict image quality in addition to other labels to increase model stability. The model was tuned to exclude images that were of very poor quality (e.g., completely over- or underexposed). Examples of excluded images are in Supplementary Figure S1. Aggregate analyses of the excluded images are in Supplementary Table S1. We preprocessed the images for training and validation and trained the neural network following the same procedure as in Gulshan et al.[11] We trained separate models to predict spherical power, cylindrical power, and SE (Fig. 1).

We used an early stopping criteria[24] based on performance on the tuning data set to help avoid overfitting and to terminate training when the model performance, such as mean absolute error (MAE), on a tuning data set stopped improving. To further improve results, we averaged the results of 10 neural network models that were trained on the same data (ensembling).[25]

Network Architecture

Our network consists of a residual network[16] (ResNet) to learn predictive image features, a soft-attention layer to select and understand the most important image features, and fully connected layers to learn interactions between the selected features.

Specifically, the residual network consists of one convolutional layer, followed by three residual blocks with four, five, and two residual units, respectively. Each residual unit has a bottleneck architecture composed of three convolutional layers. Such a residual architecture enables deeper networks (in our case 34 layers), thereby learning more abstract features and having higher prediction accuracy.[16] Of importance, skip-connections enable the network to bypass certain residual units and reuse features from different abstraction levels and resolutions, which enables the following attention layer to access more precisely localized predictive image features. The output of the residual network is a convolutional feature map A, with $A_{i,j}$ being the learned features for each spatial location (i, j).

The following soft-attention layer predicts for each location (i, j) scalar weights $w_{i,j}$, which indicate the importance of certain image regions and thereby enable understanding which image features, for example, the fovea in retina images, are most predictive. We generated individual attention maps by visualizing the predicted feature weights $w_{i,j}$ as a heat map. We also generated aggregated attention maps by averaging predicted attention weights over multiple images. The output of the soft-attention layer is a single feature vector obtained by averaging $A_{i,j}$ weighted by $w_{i,j}$ The soft-attention layer is followed by two fully connected layers and an output layer, which predicts the SE, as well as the spherical and cylindrical component (Fig. 1).

Evaluating the Algorithm

We optimized for minimizing the MAE to evaluate model performance for predicting refractive error. We also calculated the R^2 value, but this was not used to select the operating points for model performance. In addition, to further characterize the performance of the algorithms, we examined how frequently the algorithms' predictions fell within a given error margin (see Statistical Analysis section).

Statistical Analysis

To assess the statistical significance of these results, we used the nonparametric bootstrap procedure: from the validation set of N instances, we sampled N instances with replacement and evaluated the model on this sample. By repeating this sampling and evaluation 2000 times, we obtained a distribution of the performance metric (e.g., MAE) and reported the 2.5 and 97.5 percentiles as 95% confidence intervals (CIs). We compared the algorithms' MAE to baseline accuracy, which was generated by calculating the MAE of the actual refractive error and the average refractive error.

To further assess statistical significance, we performed hypothesis testing using a 1-tailed binomial test for the

Table 2. MAE and Coefficient of Determination (R^2) of Algorithm Versus Baseline for Predicting the SE

	MAE		R^2	
Data Set	**Model**	**Baseline**	**Model**	**Baseline**
UK Biobank, $n = 23{,}520$, 95% CI	0.56 [0.55, 0.56]	1.81 [1.79–1.84]	0.90 [0.90, 0.91]	0.0 [0.0, 0.0]
AREDS, $n = 7{,}635$ 95% CI	0.91 [0.89, 0.93]	1.63 [1.60–1.67]	0.69 [0.66, 0.71]	0.0 [0.0, 0.0]

Baseline metrics are calculated by predicting mean values of the validations set. All the values are in units of diopters.

frequency of the model's prediction lying within several error margins for each prediction. The baseline accuracy (corresponding to the null hypothesis) was obtained by sliding a window of size equal to the error bounds (e.g., size 1 for ±0.5) across the population histogram and taking the maximum of the summed histogram counts. This provided the maximum possible random accuracy (by guessing the center of the sliding window containing the maximum probability mass).

Attention Maps

To visualize the most predictive eye features, we integrated a soft-attention layer into our network architecture. The layer takes as input image features learned by the preceding layers, predicts for each feature a weight that indicates its importance for making a prediction, and outputs the weighted average of image features. We generated individual attention maps of images by visualizing the predicted feature weights as a heat map. We also generated aggregated attention maps by averaging predicted attention weights over multiple images.

Results

The baseline characteristics of the UK Biobank and AREDS cohorts are summarized in Table 1. Participants in the UK Biobank data set were imaged once. Subjects in the AREDS data set were imaged multiple times during the course of the trial. The subjects in the AREDS study were on average older than those in UK Biobank (mean age: 73.8 years in AREDS versus 56.8 years in UK Biobank). Hypermetropia was more common in the AREDS data set. The distribution of sex and ethnicity were similar in the two groups.

Table 2 summarizes the performance of the model on the clinical validation sets from UK Biobank and AREDS. The model was trained jointly on both the UK Biobank and AREDS data sets to predict the SE of the refractive error. Both UK Biobank and AREDS data sets reported SE, but the individual spherical and cylindrical components were available only in the UK Biobank data set. The MAE of the model on UK Biobank clinical validation data was 0.56 diopter (D) (95% CI: 0.55–0.56) and 0.91 D (95% CI: 0.89–0.93) on the AREDS clinical validation data set (see Table 2). The distribution of the predicted versus actual values for both data sets are visualized in Figure 2. The model's predicted values were within 1 D of the actual values 86% of the time for the UK Biobank clinical validation set versus 50% for baseline accuracy. For AREDS, the model's prediction was within 1 D 65% of the time versus 45% for baseline. The difference between the model and baseline were significant at all margins of error (Supplementary Table S1).

We further trained separate models to predict the components of SE, spherical power, and cylindrical power, using the UK Biobank data set because these values were not available in the AREDS data set. The model trained to predict the spherical component from retinal fundus images was quite accurate, with an MAE of 0.63 D (95% CI: 0.63, 0.64), and R^2 of 0.88 (95% CI: 0.88, 0.89). In comparison, the model trained to predict cylindrical power was not very accurate, with an MAE of 0.43 (95% CI: 0.42, 0.43) and R^2 of 0.05 (95% CI: 0.04, 0.05) (see Supplementary Table S2).

Attention maps were generated to visualize the regions on the fundus that were most important for the refractive error prediction. Representative examples of attention maps at different categories of severities of refractive error (myopia, hyperopia) are shown in Figure 3. For every image, the macula was a prominent feature that was highlighted. In addition, diffuse signals such as retinal vessels and cracks in retinal pigment were also highlighted. There was not an obvious difference in the heat maps for different severities of refractive

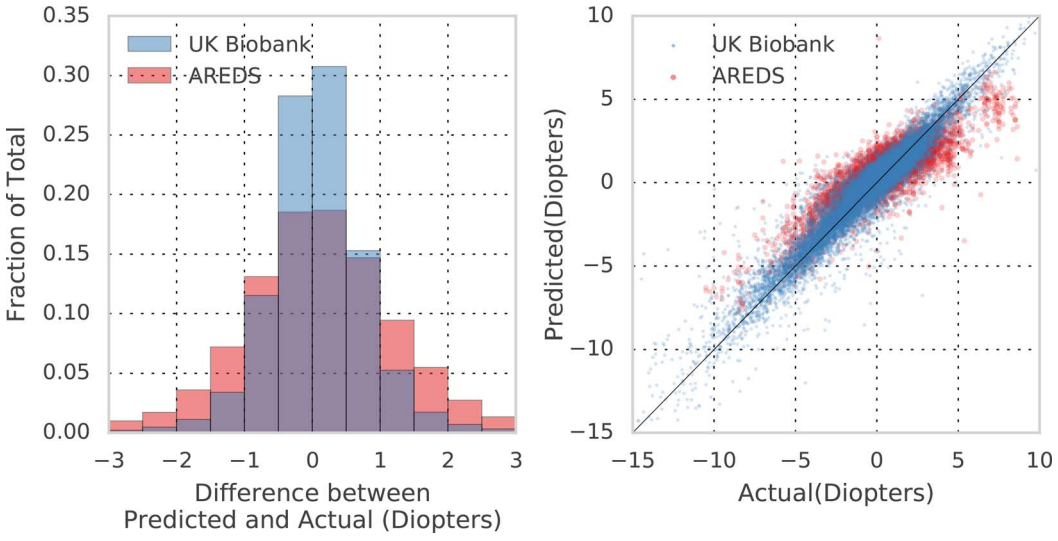

Figure 2. Model performance in predicting SE on the two clinical validation sets. (**A**) Histogram of prediction error (Predicted − Actual) UK Biobank data set (*blue*) and AREDS data set (*red*). (**B**) Scatter plot of predicted and actual values for each instance in the validation sets. *Black diagonal* indicates perfect prediction, where $y = x$.

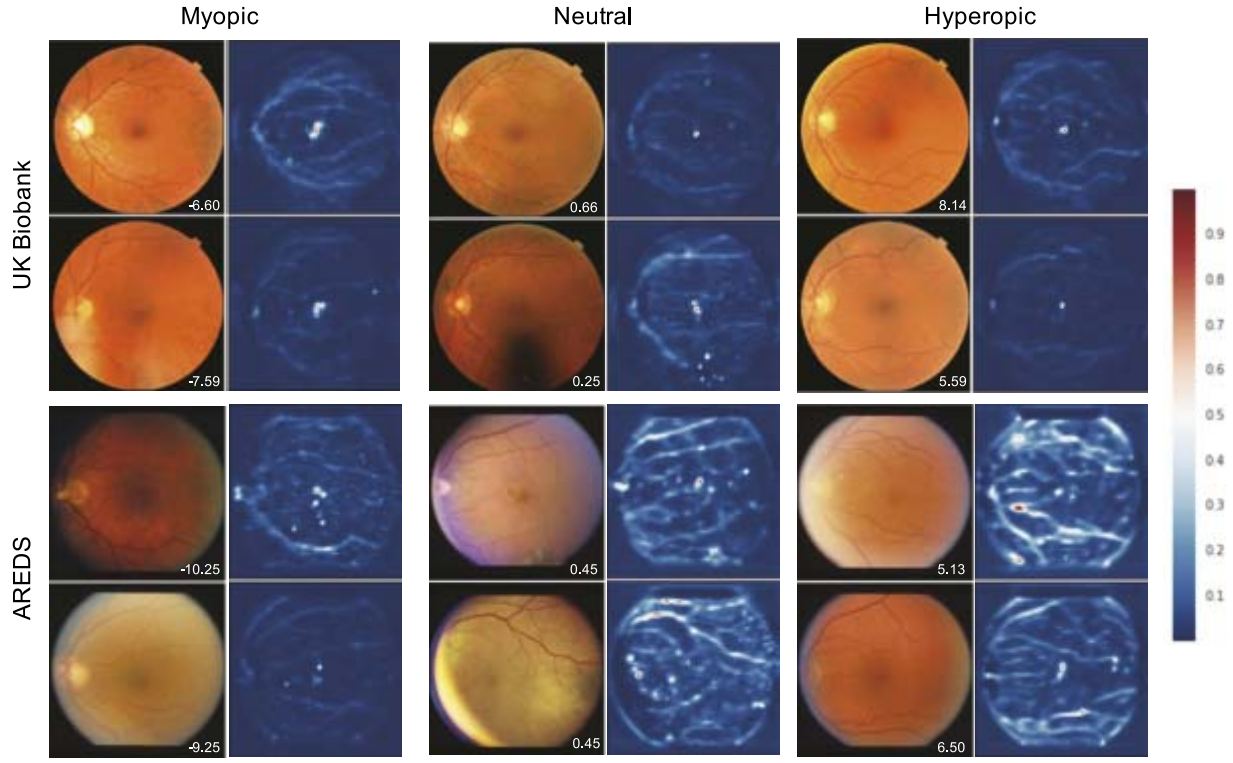

FIGURE 3. Example attention maps for three left myopic (SE worse than −6.0), neutral (SE between −1.0 and 1.0), and hyperopic (SE worse than 5.0) fundus images from UK Biobank (*two top rows*) and AREDS (*two bottom rows*). Diagnosed SE is printed in the bottom right corner of fundus images. Scale bar on *right* denotes attention pixel values, which are between 0 and 1 (exclusive), with the sum of all values equal to 1.

error. We averaged and merged the attention maps for 1000 images at different severities of refractive error and found that these observations also generalized across many images (Supplementary Figs. S2 and S3). To ensure that the attention heat maps involved the fovea and not simply the center of the image, we also automatically aligned the images on the fovea and attained the same result (Supplementary Fig. S4). Given the importance of the fovea in the model predictions, we also investigated the effect that eye disease may have on the accuracy of predictions. The UK Biobank data set contained mostly healthy eyes and so could not be used for this analysis. Using the AREDS data set, we subdivided the subject population based upon whether or not the subject had cataract surgery and/or AMD. We found a small but significant improvement in the accuracy of the model when we excluded subjects who had cataract surgery and/or AMD from the analysis (Supplementary Table S3).

DISCUSSION

In this study, we have shown that deep learning models can be trained to predict refractive error from retinal fundus images with high accuracy, a surprising result given that this was not a prediction task thought to be possible from retinal fundus images. Both individual and mean attention maps, which highlight features predictive for refractive error, show a clear focus of attention to the fovea for all refractive errors. While attention maps show anatomical correlates for the prediction of interest, they do not establish causation. This is a general limitation of existing attention techniques. In addition, we averaged a large set of attention maps to examine them in aggregate. Because it is possible that predictive anatomical features might vary in their location in the images, it is possible these activations averaged out in the mean attention map. However, these maps may be a way to generate hypotheses in a nonbiased manner to further research into the pathophysiology of refractive error.

For example, the consistent focus on the fovea shown in the attention maps may be an avenue for further myopia research. Given that fundus images are generally centered on the fovea, perhaps this result is associated with the spatial relationship between the fovea and the other retinal landmarks. However, it is also possible that the appearance of the fovea itself holds information about refractive error. In pathological myopia, the fundus may display characteristic clinical signs that involve the macula.[26] However, to the best of our knowledge, other than in pathological myopia, there is no prior literature exploring the relationship between the foveal architecture imaged using a fundus camera and refractive error or axial length. Previous work with higher resolution using OCT has shown some evidence for anatomical difference in the retinal thickness or contour at the fovea with varying refractive error.[27] Although there is some evidence for greater spacing of foveal cone photoreceptors in myopic eyes,[28] this is unlikely to be resolved in retinal fundus images. One hypothesis may be that there is a variation in the reflectance or focus of the fovea with varying refractive error when imaged using a fundus camera. When visualized using an ophthalmoscope, the foveal light reflex becomes dimmer and less conspicuous with increasing age[29] or presence of macular disease. However, the "brightness" of this reflex and its relationship with refractive error has not been studied. Another hypothesis may be that there is a relationship between color or macular pigment at the fovea and refractive error; however Czepita et al.[30] found no association. The density of pigment is usually derived using psychophysical techniques, but fundus photographs captured using blue and green illumination have shown promise in evaluating density.[31]

The attention maps also suggest that features outside the foveal region contribute to the prediction to a lesser extent, including a diffuse signal from the optic nerve head (ONH) and retinal temporal vessel arcades from their exit from the optic nerve as they traverse across the fundus. The extent of association between optic disc size and refractive error is unresolved due to inconsistent findings among studies. Some studies have shown a weakly significant increase in optic disc size with increasing refractive error toward myopia,[32,33] whereas a Chinese population-based study found that the optic disc size is independent of refractive error within the range of −8 to +4 D.[34] Varma et al.[35] found no association between refractive error and optic disc size. Beyond size, the appearance of the optic disc may vary with refractive error, and eyes with axial myopia may display tilted optic discs.[36] Myopic refractive errors have also been associated with narrower retinal arterioles and venules and increased branching,[37] and reduction in retinal vascular fractal dimensions.[38] In addition, the maps also looked very similar for images with hypermetropia and myopia, suggesting that the neural network is leveraging the same regions for predictions over a spectrum of refractive errors.

While each landmark may individually contribute to the prediction, the relationship between retinal landmarks could be equally the predictive feature. The spatial relationships between anatomical features in relation to ametropia have been studied extensively.[39–41] As the most predictive features of the model are the fovea and ONH, a spatial relationship between these two points, as well as other landmarks, should be considered. Baniasadi et al.[42] found that a combination of parameters related to the ONH, namely the interartery angle between superior and inferior temporal arteries, ONH tilt and rotation, and the location of the central retinal vessel trunk had a strong association with SE. Although both the ONH and vessels are clearly highlighted in both the averaged attention maps, the strength of attention to these regions is difficult to determine in the averaged attention maps as the signals from the ONH and retinal vessels were much more diffuse due to the interindividual variance of their locations. Additionally, the difference in this signal between myopic and hypermetropic eyes is not immediately discernible. Analyzing the attention maps and the spatial relationships between the predictive regions at eye level would be an area for further study.

We found that the MAE of our joint model on the Biobank data set was lower than on the AREDS data set and that there potentially may be greater error in the SE prediction at the extremes of refractive error. These observations may be due to a variety of factors. Firstly, the camera used to image the fundus in the UK Biobank study was a wider 45-degree field camera that captured more peripheral information than did the 30-degree field of the Zeiss camera used in the AREDS data set. This results in the optic disc or retinal vessels (shown to be important in the UK Biobank model) not always being visible in the acquired image. Secondly, the AREDS data set has far fewer images, and small training sets generally result in a decrease in generalizability and performance in the clinical validation set. In addition, many images in the AREDS data set exhibited macular pathology of some form that may add noise to the images, resulting in more widespread areas of attention (Supplementary Fig. S3). Given the importance of the foveal region in prediction refractive error, this might have decreased the model's performance on the AREDS data set. Thirdly, the refractive error was determined by two different methods in each data set: autorefraction in the Biobank data set versus subjective refraction in AREDS. We believe that the smaller capture field, preexisting eye pathologies, and smaller data set combined lowered the predictive power in the AREDS data set relative to the UK Biobank. However, future studies would be required to test and quantify the influence that each of these factors has on model accuracy.

The model has high accuracy when predicting spherical power but not when predicting cylindrical power. This is expected as astigmatism is the result of toricity of the cornea and/or the crystalline lens, information that is unlikely to be held in retinal fundus images. As described earlier, retinal features associated with refractive errors may be related to differing axial lengths. Therefore, the high accuracy of prediction of SE is likely predicting axial ametropia. Spherical power related to lens ametropia is not known to have any specific relationship with retinal anatomy. However, lens phenomena, such as the hypermetropic shift, that is, the increasing thickness of the crystalline lens due to aging, may affect the focus settings of the camera and consequently result in magnification effects of the image. Wang et al.[43] suggested that focus and magnification effects are age dependent after ages above 42 years, related to presbyopic changes in the lens. We found that the predicted SE was slightly underestimated in the AREDS group, particularly in the hypermetropic eyes. This group was significantly older (approximately 20 years older on average) than the UK Biobank group and therefore may have experienced presbyopia with a hypermetropic shift. As this is lens ametropia as opposed to axial ametropia, the model was unable to identify this. Unfortunately, axial length data was unavailable for either data set to investigate the hypothesis that its relationship with spherical refractive error is the source of the prediction. Future studies with a data set that includes axial length would help elucidate this question.

Additional future work should include data sets from even more diverse populations, such as different ethnicities, ages, and comorbidities. The model was trained and validated on a combination of two data sets. It would be more desirable to have a third data set that was taken in a completely different setting for additional validation. In addition, the UK Biobank data set excluded patients who had prior eye surgery. Additional work could include adding these patients back in to the data to see its effects on model performance.

Portable fundus cameras such as PEEK[44] are becoming less expensive and more common for screening and diagnosis of eye disease, particularly in the developing world. With further validation, it may be possible to use these increasingly abundant fundus images to efficiently screen for individuals with uncorrected refractive error who would benefit from a formal refraction assessment. However, currently, autorefraction is no more difficult to perform than fundus photography, so the findings of this study are unlikely to change the role of autorefraction in most clinical settings.

Nevertheless, the methods and results of this article represent new approaches to biological and ophthalmologic research. The development of highly accurate automated classification algorithms can aid in research that involves large-scale retrospective data sets. For example, this algorithm could help in epidemiologic research of myopia from large fundus image data sets that do not have refractive error labels. The attention map results produced by this study may aid in deeper understanding of the biology and pathophysiology of myopia. Lastly, the process used in this study—leveraging deep learning to first directly predict the outcome or phenotype of interest and then attention techniques to localize the most predictive features—could be a method that can be applied to catalyze scientific research broadly in medicine and biology.

Acknowledgments

The authors thank Mark DePristo, PhD, Arunachalam Narayanaswamy, PhD, and Yun Liu, PhD, from Google Research for their technical advice and review of the manuscript.

Supported by UK Biobank Resource under Application Number 17643.

Disclosure: **A.V. Varadarajan**, Google LLC (E); **R. Poplin**, Google LLC (E); **K. Blumer**, Google LLC (E); **C. Angermueller**, Google LLC (E); **J. Ledsam**, Google LLC (E); **R. Chopra**, Google LLC (C); **P.A. Keane**, Google LLC (C); **G.S. Corrado**, Google LLC (E); **L. Peng**, Google LLC (E); **D.R. Webster**, Google LLC (E)

References

1. Pascolini D, Mariotti SP. Global estimates of visual impairment: 2010. *Br J Ophthalmol*. 2012;96:614–618.
2. Rabbetts R. *Bennett and Rabbett's Clinical Visual Optics*. 4th ed. Amsterdam: Elsevier; 2007.
3. Foster PJ, Jiang Y. Epidemiology of myopia. *Eye*. 2014;28:202–208.
4. Kovin S, Naidoo JJ. Uncorrected refractive errors. *Indian J Ophthalmol*. 2012;60:432–437.
5. Bolster NM, Giardini ME, Livingstone IAT, Bastawrous A. How the smartphone is driving the eye-health imaging revolution. *Expert Rev Ophthalmol*. 2014;9:475–485.
6. Bastawrous A, Rono HK, Livingstone IAT, et al. Development and validation of a smartphone-based visual acuity test (peek acuity) for clinical practice and community-based fieldwork. *JAMA Ophthalmol*. 2015;133:930–937.
7. Bolster NM, Giardini ME, Bastawrous A. The diabetic retinopathy screening workflow: potential for smartphone imaging. *J Diabetes Sci Technol*. 2015;10:318–324.
8. Bastawrous A, Giardini ME, Bolster NM, et al. Clinical validation of a smartphone-based adapter for optic disc imaging in Kenya. *JAMA Ophthalmol*. 2016;134:151–158.
9. LeCun Y, Yoshua B, Geoffrey H. Deep learning. *Nature*. 2015; 521:436–444.
10. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542:115–118.
11. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316:2402–2410.
12. Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*. 2017;124: 962–969.
13. Liu Y, Gadepalli K, Norouzi M, et al. Detecting cancer metastases on gigapixel pathology images. *arXiv*. 2017: 1703.02442.
14. Wang D, Khosla A, Gargeya R, Irshad H, Beck AH. Deep learning for identifying metastatic breast cancer. *arXiv*. 2016: 1606.05718.
15. Poplin R, Varadarajan AV, Blumer K, et al. Predicting cardiovascular risk factors from retinal fundus photographs using deep learning. *arXiv*. 2017:1708.09843.
16. He K, Zhang X, Ren S, Sun J. Identity mappings in deep residual networks. *Comput Vis ECCV*. 2016;4:630–645.
17. Atchison DA, Pritchard N, Schmid KL, Scott DH, Jones CE, Pope JM. Shape of the retinal surface in emmetropia and myopia. *Invest Ophthalmol Vis Sci*. 2005;46:2698–2707.
18. Biobank. UK About UK Biobank website. Available at: http://www.ukbiobank.ac.uk/about-biobank-uk/. Accessed March 26, 2017.
19. National Eye Institute. National Eye Institute (NEI) Age-Related Eye Disease Study (AREDS) (dbGaP ID: phs000001.v3.p1). Available at: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000001.v3.p1. Accessed September 22, 2017.
20. National Eye Institute. Age-Related Eye Disease Study (AREDS). Photographic procedures. Chapter 8. dbGaP ID: phd000008. https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/document.cgi?study_id=phs000001.v3.p1&phv=53743&phd=8&pha=2856&pht=371&phvf=&phdf=&phaf=&phtf=&dssp=1&consent=&temp=1. Accessed November 22, 2017.
21. Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. *Mol Syst Biol*. 2016;12:878.
22. TensorFlow. Open source machine learning framework. http://tensorflow.org. Accessed August 3, 2017.
23. Xu K, Ba J, Kiros R, et al. Show, attend and tell: neural image caption generation with visual attention. *Proc Int Conf Mach Learn Appl*. 2015:2048–2057.
24. Caruana R, Lawrence S, Giles L. Overfitting in neural nets: backpropagation, conjugate gradient, and early stopping. In: Leen TK, Dietterich TG, Tresp V. *Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference*. Cambridge, MA: MIT Press; 2001.
25. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Adv Neural Inf Process Syst*. 2012:1097–1105.
26. Ryan S, Schachat A, Wilkinson C, Hinton D, Sadda S, Wiedemann P. *Retina*. 5th ed. Amsterdam: Elsevier; 2012.
27. Ostrin LA, Yuzuriha J, Wildsoet CF. Refractive error and ocular parameters: comparison of two SD-OCT systems. *Optom Vis Sci*. 2015;92:437–446.
28. Kitaguchi Y, Bessho K, Yamaguchi T, Nakazawa N, Mihashi T, Fujikado T. In vivo measurements of cone photoreceptor spacing in myopic eyes from images obtained by an adaptive optics fundus camera. *Jpn J Ophthalmol*. 2007;51:456–461.
29. Newcomb RD, Potter JW. Clinical investigation of the foveal light reflex. *Am J Optom Physiol Opt*. 1981;58:1110–1119.
30. Czepita M, Karczewicz D, Safranow K, Czepita D. Macular pigment optical density and ocular pulse amplitude in subjects with different axial lengths and refractive errors. *Med Sci Monit*. 2015;21:1716–1720.
31. Bour LJ, Koo L, Delori FC, Apkarian P, Fulton AB. Fundus photography for measurement of macular pigment density distribution in children. *Invest Ophthalmol Vis Sci*. 2002;43: 1450–1455.
32. Ramrattan RS, Wolfs RC, Jonas JB, Hofman A, de Jong PT. Determinants of optic disc characteristics in a general population: The Rotterdam Study. *Ophthalmology*. 1999; 106:1588–1596.
33. Wu R-Y, Wong T-Y, Zheng Y-F, et al. Influence of refractive error on optic disc topographic parameters: the Singapore Malay Eye Study. *Am J Ophthalmol*. 2011;152:81–86.
34. Jonas JB. Optic disk size correlated with refractive error. *Am J Ophthalmol*. 2005;139:346–348.
35. Varma R, Tielsch JM, Quigley HA, et al. Race-, age-, gender-, and refractive error-related differences in the normal optic disc. *Arch Ophthalmol*. 1994;112:1068–1076.
36. Vongphanit J, Mitchell P, Wang JJ. Population prevalence of tilted optic disks and the relationship of this sign to refractive error. *Am J Ophthalmol*. 2002;133:679–685.
37. Lim LS, Cheung CY-L, Lin X, Mitchell P, Wong TY, Mei-Saw S. Influence of refractive error and axial length on retinal vessel geometric characteristics. *Invest Ophthalmol Vis Sci*. 2011; 52:669–678.
38. Li H, Mitchell P, Liew G, et al. Lens opacity and refractive influences on the measurement of retinal vascular fractal dimension. *Acta Ophthalmol*. 2010;88:e234–e240.
39. Elze T, Baniasadi N, Jin Q, Wang H, Wang M. Ametropia, retinal anatomy, and OCT abnormality patterns in glaucoma. 1. Impacts of refractive error and interartery angle. *J Biomed Opt*. 2017;22:121713.
40. Wang M, Jin Q, Wang H, Li D, Baniasadi N, Elze T. The interrelationship between refractive error, blood vessel

anatomy, and glaucomatous visual field loss. *Trans Vis Sci Tech*. 2018;7(1):4.

41. Wang M, Elze T, Li D, et al. Age, ocular magnification, and circumpapillary retinal nerve fiber layer thickness. *J Biomed Opt*. 2017;22:1–19.
42. Baniasadi N, Wang M, Wang H, Mahd M, Elze T. Associations between optic nerve head–related anatomical parameters and refractive error over the full range of glaucoma severity. *Trans Vis Sci Tech*. 2017;6(4):9.
43. Wang M, Elze T, Li D, et al. Age, ocular magnification, and circumpapillary retinal nerve fiber layer thickness. *J Biomed Opt*. 2017;22:1–19.
44. Peek Vision. Peek Vision: Vision & Health for Everyone website. Available at: http://10ga.iapb.org/peek-vision-vision-health-for-everyone. Accessed December 4, 2017.

Investigative Ophthalmology & Visual Science