https://doi.org/10.1038/s41551-018-0195-0

Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning

Ryan Poplin[1,4], Avinash V. Varadarajan[1,4], Katy Blumer[1], Yun Liu[1], Michael V. McConnell[2,3], Greg S. Corrado[1], Lily Peng[1,4]* and Dale R. Webster[1,4]

Traditionally, medical discoveries are made by observing associations, making hypotheses from them and then designing and running experiments to test the hypotheses. However, with medical images, observing and quantifying associations can often be difficult because of the wide variety of features, patterns, colours, values and shapes that are present in real data. Here, we show that deep learning can extract new knowledge from retinal fundus images. Using deep-learning models trained on data from 284,335 patients and validated on two independent datasets of 12,026 and 999 patients, we predicted cardiovascular risk factors not previously thought to be present or quantifiable in retinal images, such as age (mean absolute error within 3.26 years), gender (area under the receiver operating characteristic curve (AUC) = 0.97), smoking status (AUC = 0.71), systolic blood pressure (mean absolute error within 11.23 mmHg) and major adverse cardiac events (AUC = 0.70). We also show that the trained deep-learning models used anatomical features, such as the optic disc or blood vessels, to generate each prediction.

Risk stratification is central to identifying and managing groups at risk for cardiovascular disease, which remains the leading cause of death globally[1]. Although the availability of cardiovascular disease risk calculators, such as the Pooled Cohort equations[2], Framingham[3–5] and Systematic COronary Risk Evaluation (SCORE)[6,7], is widespread, there are many efforts to improve risk predictions. Phenotypic information, particularly of vascular health, may further refine or reclassify risk prediction on an individual basis. Coronary artery calcium is one such example, for which it has been shown that additional signals from imaging improve risk stratification[8]. The current standard-of-care for the screening of cardiovascular disease risk requires a variety of variables derived from the patient's history and blood samples, such as age, gender, smoking status, blood pressure, body mass index (BMI), glucose and cholesterol levels[9]. Most cardiovascular risk calculators use some combination of these parameters to identify patients at risk of experiencing either a major cardiovascular event or cardiac-related mortality within a pre-specified time period, such as ten years. However, some of these parameters may be unavailable. For example, in a study from the Practice INNovation And CLinical Excellence (PINNACLE) electronic-health-record-based cardiovascular registry, the data required to calculate the 10-year risk scores were available for less than 30% of the patients[10]. This was largely due to missing cholesterol values[10], which is not surprising given that a fasting blood draw is required to obtain these data. In this situation, BMI can be used in the place of lipids for a preliminary assessment of cardiovascular health[11–13]. We therefore explored whether additional signals for cardiovascular risk can be extracted from retinal images, which can be obtained quickly, cheaply and non-invasively in an outpatient setting.

Markers of cardiovascular disease, such as hypertensive retinopathy and cholesterol emboli, can often manifest in the eye. Furthermore, because blood vessels can be non-invasively visualized from retinal fundus images, various features in the retina, such as vessel calibre[14–20], bifurcation or tortuosity[21], microvascular changes[22,23] and vascular fractal dimensions[24–26], may reflect the systemic health of the cardiovascular system as well as future risk. The clinical utility of such features still requires further study. In this work, we demonstrate the extraction and quantification of multiple cardiovascular risk factors from retinal images using deep learning.

Machine learning has been leveraged for many years for a variety of classification tasks, including the automated classification of eye disease. However, much of the work has focused on 'feature engineering', which involves computing explicit features specified by experts[27,28]. Deep learning is a family of machine-learning techniques characterized by multiple computation layers that allow an algorithm to learn the appropriate predictive features on the basis of examples rather than requiring features to be hand-engineered[29]. Recently, deep convolutional neural networks—a special type of deep-learning technique that has been optimized for images—have been applied to produce highly accurate algorithms that diagnose diseases, such as melanoma[30] and diabetic retinopathy[31,32], from medical images, with comparable accuracy to that of human experts.

Results

We developed deep-learning models using retinal fundus images from 48,101 patients from the UK Biobank (http://www.ukbiobank.ac.uk/about-biobank-uk) and 236,234 patients from EyePACS (http://www.eyepacs.org) and validated these models using images from 12,026 patients from the UK Biobank and 999 patients from EyePACS (Table 1). The mean age was 56.9 ± 8.2 years on the UK Biobank clinical validation dataset and 54.9 ± 10.9 years in the EyePACS-2K clinical validation dataset. The UK Biobank population was predominantly Caucasian, while the EyePACS patients were predominantly Hispanic. Haemoglobin A1c (HbA1c) measurements were available only in 60% of the EyePACS population. Because this population consisted of mostly diabetic patients presenting for diabetic retinopathy screening, the mean HbA1c level of this population was 8.2 ± 2.1%—well above the normal range. UK Biobank participants were recruited from a UK general population,

[1]Google Research, Google, Mountain View, CA, USA. [2]Verily Life Sciences, South San Francisco, CA, USA. [3]Division of Cardiovascular Medicine, Stanford School of Medicine, Stanford, CA, USA. [4]These authors contributed equally: Ryan Poplin, Avinash V. Varadarajan, Lily Peng and Dale R. Webster. *e-mail: lhpeng@google.com

Table 1 | Baseline characteristics of patients in the development and validation sets

Characteristics	Development set		Clinical validation set	
	UK Biobank	EyePACS	UK Biobank	EyePACS-2K
Number of patients	48,101	236,234	12,026	999
Number of images	96,082	1,682,938	24,008	1,958
Age: mean, years (s.d.)	56.8 (8.2), *n* = 48,101	53.6 (11.6), *n* = 234,140	56.9 (8.2), *n* = 12,026	54.9 (10.9), *n* = 998
Gender (% male)	44.9, *n* = 48,101	39.2, *n* = 236,212	44.9, *n* = 12,026	39.2, *n* = 999
Ethnicity	1.2% Black, 3.4% Asian/PI, 90.6% White, 4.1% Other *n* = 47,785	4.9% Black, 5.5% Asian/PI, 7.7% White, 58.1% Hispanic, 1.2% Native American, 1.7% Other *n* = 186,816	1.3% Black, 3.6% Asian/PI, 90.1% White, 4.2% Other *n* = 11,926	6.4% Black, 5.7% Asian/PI, 11.3% White, 57.2% Hispanic, 0.7% Native American, 2% Other *n* = 832
BMI: mean (s.d.)	27.31 (4.78), *n* = 47,847	n/a	27.37 (4.79), *n* = 11,966	n/a
Systolic BP: Mean, mmHg (s.d.)	136.82 (18.41), *n* = 47,918	n/a	136.89 (18.3), *n* = 11,990	n/a
Diastolic BP: Mean, mmHg (s.d.)	81.78 (10.08), *n* = 47,918	n/a	81.76 (9.87), *n* = 11,990	n/a
HbA1c: mean, % (s.d.)	n/a	8.23 (2.14), *n* = 141,715	n/a	8.2 (2.13), *n* = 737
Current Smoker: %	9.53%, *n* = 47,942	n/a	9.87%, *n* = 11,990	n/a

n/a indicates that the characteristic was not available for that dataset. *n* is the number of patients for whom that measurement was available. PI, Pacific Islander.

rather than a diabetic population presenting for diabetic retinopathy screening. Fasting glucose, HbA1c levels and other laboratory-based methods for determining diabetes status were not available at the time of this study. However, approximately 5% of the UK Biobank population self-identified as having 'diabetes diagnosed by doctor' (without further specification of type). Additional patient demographics are summarized in Table 1.

First, we tested the ability of our models to predict a variety of cardiovascular risk factors from retinal fundus images (Tables 2 and 3). Because of the lack of an established baseline for predicting these features from retinal images, we use the average value as the baseline for continuous predictions (such as age). The mean absolute error (MAE) for predicting the patient's age was 3.26 years (95% confidence interval (CI): 3.22 to 3.31) versus baseline (7.06; 95% CI: 6.98 to 7.13) in the UK Biobank validation dataset. For the EyePACS-2K, the MAE was 3.42 (95% CI: 3.23 to 3.61) versus baseline (8.48; 95% CI: 8.07 to 8.90). The predicted age and actual age have a fairly linear relationship (Fig. 1a), which is consistent over both datasets. The algorithms also predicted systolic blood pressure (SBP), BMI and HbA1c better than baseline (Table 2). However, these predictions had low coefficient of determination (R^2) values, suggesting that—although better than baseline—the algorithm is not able to predict these parameters with high precision. The predicted SBP increased linearly with actual SBP until approximately 150 mmHg, but levelled off above that value (Fig. 1b).

To further characterize the performance of the algorithms, we examined how frequently the algorithms' predictions fell within a given error margin and compared this with the baseline accuracy (Table 3). Although the models were not optimized for this task (for instance, for age they were optimized to minimize MAE rather than to predict age within specific error margins), we found that the algorithm performed significantly better than baseline for age, SBP, diastolic blood pressure (DBP) and BMI. For example, we found that in 78% of the cases the predicted age was within a ±5-year

Table 2 | Algorithm performance on predicting cardiovascular risk factors in the two validation sets

Predicted risk factor (evaluation metric)	UK Biobank validation dataset (*n* = 12,026 patients)		EyePACS-2K validation dataset (*n* = 999 patients)	
	Algorithm (95% CI)	Baseline	Algorithm (95% CI)	Baseline
Age: MAE, years (95% CI)	3.26 (3.22,3.31)	7.06 (6.98,7.13)	3.42 (3.23,3.61)	8.48 (8.07,8.90)
Age: R^2 (95% CI)	0.74 (0.73,0.75)	0.00	0.82 (0.79,0.84)	0.00
Gender: AUC (95% CI)	0.97 (0.966,0.971)	0.50	0.97 (0.96,0.98)	0.50
Current smoker: AUC (95% CI)	0.71 (0.70,0.73)	0.50	n/a	n/a
HbA1c: MAE, % (95% CI)	n/a	n/a	1.39 (1.29,1.50)	1.67 (1.58,1.77)
HbA1c: R^2 (95% CI)	n/a	n/a	0.09 (0.03,0.16)	0.00
SBP: MAE, mmHg (95% CI)	11.35 (11.18,11.51)	14.57 (14.38,14.77)	n/a	n/a
SBP: R^2 (95% CI)	0.36 (0.35,0.37)	0.00	n/a	n/a
DBP: MAE, mmHg (95% CI)	6.42 (6.33,6.52)	7.83 (7.73,7.94)	n/a	n/a
DBP: R^2 (95% CI)	0.32 (0.30,0.33)	0.00	n/a	n/a
BMI: MAE (95% CI)	3.29 (3.24,3.34)	3.62 (3.57,3.68)	n/a	n/a
BMI: R^2 (95% CI)	0.13 (0.11,0.14)	0.00	n/a	n/a

95% CIs on the metrics were calculated with 2,000 bootstrap samples (Methods). For continuous risk factors (such as age), the baseline value is the MAE of predicting the mean value for all patients.

Table 3 | Model accuracy versus baseline for predicting various continuous risk factors within a given margin

Predicted risk factor	UK Biobank validation dataset (*n* = 12,026 patients)				EyePACS-2K validation dataset (*n* = 999 patients)			
	Error margin	Model accuracy (%)	Baseline accuracy (%)[a]	*P* value	Error margin	Model accuracy (%)	Baseline accuracy (%)[a]	*P* value
Age (years)	±1	20	11	<0.0001	±1	20%	13%	<0.0001
	±3	54	29	<0.0001	±3	56%	28%	<0.0001
	±5	78	44	<0.0001	±5	79%	43%	<0.0001
SBP (mmHg)	±5	29	22	<0.0001				
	±10	53	43	<0.0001				
	±15	72	60	<0.0001				
DBP (mmHg)	±3	30	25	<0.0001				
	±5	46	41	<0.0001				
	±10	79	71	<0.0001				
BMI	±1	21	20	0.02				
	±3	57	54	<0.0001				
	±5	80	77	<0.0001				
HbA1c (%)					±0.5	31	35	0.995
					±1	54	54	0.486
					±2	79	78	0.255

[a] Baseline accuracy was generated by sliding a window with a size equal to the error bounds (for example, size 10 for ±5) across the population histogram and then taking the maximum of the summed histogram counts. This provides the maximum possible 'random' accuracy (by guessing the centre of the sliding window corresponding to the maximum). *P* values were obtained using a one-tailed binomial test with *n* = number of patients, with the baseline accuracy as the chance probability of a correct prediction.

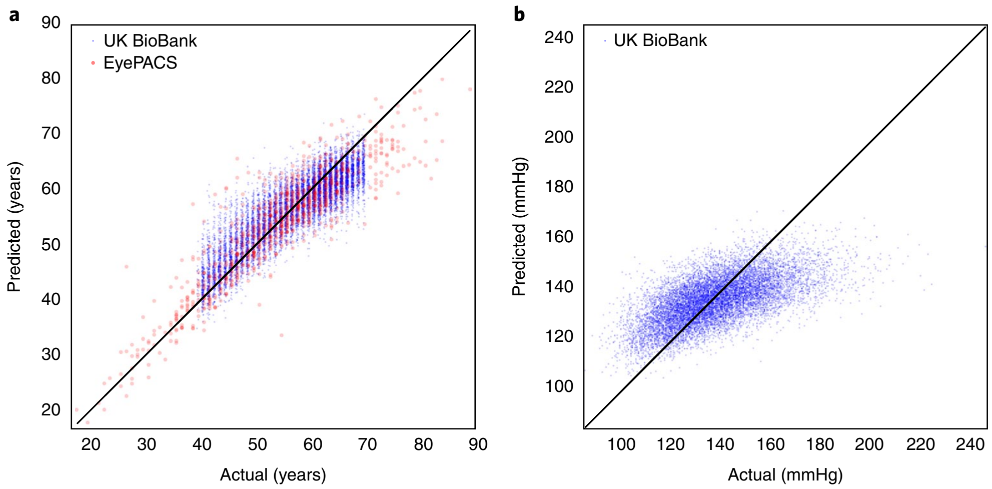

Fig. 1 | Predictions of age and SBP. a, Predicted and actual age in the two validation datasets. For the UK Biobank dataset, age was calculated using the birth year because birth months and days were not available. In the EyePACS-2K dataset, age is available only in units of whole years. **b**, Predicted and actual SBP in the UK Biobank validation dataset. The lines represent *y* = *x* values.

margin of the actual age, whereas baseline predictions only fell into the 5-year window 44% of the time. Notably, the model could not predict HbA1c at an accuracy that was better than baseline. We also found that our approach was able to infer ethnicity, which is another potential cardiovascular risk factor[2] (κ score of 0.60 (95% CI: 0.58 to 0.63) in the UK Biobank validation dataset and 0.75 (95% CI: 0.70 to 0.79) in the EyePACS-2K validation dataset).

We also examined the effect that retinal eye disease such as diabetic retinopathy may have on the performance of the algorithms using the EyePACS-2K dataset, which has diabetic retinopathy grades that have been adjudicated by retinal specialists in a process previously described[33]. We stratified the model's performance by diabetic retinopathy severity and found no significant difference between the groups (Table 4).

Because retinal images alone were sufficient to predict several cardiovascular risk factors to varying degrees, we reasoned that the images could be correlated directly with cardiovascular events. Therefore, we trained a model to predict the onset of major adverse cardiovascular events (MACE) within five years. This outcome was available only for one of our datasets—the UK Biobank. Because data in the UK Biobank come from a fairly recent study that recruited relatively healthy participants, MACE were rare (631 events occurred within 5 years of retinal imaging, 150 of which were in the clinical validation dataset). Despite the limited number of

Table 4 | Performance of the algorithm stratified by the severity of diabetic retinopathy according to the international clinical diabetic retinopathy scale

	No DR	Mild DR	Moderate DR	Severe DR	Proliferative DR
Number of patients	734	81	101	32	11
Age: MAE(95% CI)	3.27 (3.06,3.49)	2.88 (2.31,3.50)	3.41 (2.82,4.06)	3.06 (2.42,3.75)	5.50 (2.64,8.92)
Age: R^2(95% CI)	0.84 (0.81,0.87)	0.87 (0.80,0.92)	0.73 (0.58,0.82)	0.85 (0.69,0.91)	0.51 (−0.45,0.92)
Gender: AUC (95% CI)	0.98 (0.97,0.99)	0.95 (0.89,0.99)	0.95 (0.90,0.99)	0.97 (0.90,1.00)	1.00 (NA)
HbA1c: MAE (95% CI)	1.28 (1.17,1.41)	1.63 (1.19,2.11)	1.77 (1.52,2.06)	1.59 (1.23,1.99)	1.33 (0.51,2.30)
HbA1c: R^2(95% CI)	0.02 (−0.04,0.08)	−0.02 (−0.31,0.20)	−0.08 (−0.36,0.13)	0.23 (−0.28,0.45)	0.51 (−3.36,0.65)

MAE and R^2 for age, gender and HbA1c are stratified by diabetic retinopathy (DR) status for the EyePACS-2K validation dataset. Categories such as severe and proliferative DR have very few patients, leading to wide CIs. NA, not applicable.

events, our model achieved an area under the receiver operating characteristic curve (AUC) of 0.70 (95% CI: 0.648 to 0.740) from retinal fundus images alone, comparable to an AUC of 0.72 (95% CI: 0.67 to 0.76) for the composite European SCORE risk calculator (Table 5). As a comparison, AUCs were also generated using individual risk factors alone. Not surprisingly, the combination of these risk factors was better at predicting MACE than individual risk factors alone.

Next, we used soft attention (Methods) to identify the anatomical regions that the algorithm might have been using to make its predictions. A representative example of a single retinal fundus image with accompanying attention maps (also called saliency maps[34]) for each prediction is shown in Fig. 2. In addition, for each prediction task, ophthalmologists blinded to the prediction task of the model assessed 100 randomly chosen retinal images to identify patterns in the anatomical locations highlighted by the attention maps for each prediction (Table 6). Encouragingly, the blood vessels were highlighted in the models trained to predict risk factors, such as age, smoking and SBP. Models trained to predict HbA1c tended to highlight the perivascular surroundings. Models trained to predict gender primarily highlighted the optic disc, vessels and macula, although there appeared to be signal distributed throughout the retina as well. For other predictions, such as SBP and BMI, the attention masks were non-specific, such as uniform 'attention' or highlighting the circular border of the image, suggesting that the signals for those predictions may be distributed more diffusely throughout the image.

Table 5 | Predicting five-year MACE in the UK Biobank validation dataset using various input variables

Risk factor(s) or model used for the prediction	AUC (95% CI)
Age only	0.66 (0.61,0.71)
SBP only	0.66 (0.61,0.71)
BMI only	0.62 (0.56,0.67)
Gender only	0.57 (0.53,0.62)
Current smoker only	0.55 (0.52,0.59)
Algorithm only	0.70 (0.65,0.74)
Age + SBP + BMI + gender + current smoker	0.72 (0.68,0.76)
Algorithm + age + SBP + BMI + gender + current smoker	0.73 (0.69,0.77)
SCORE[6,7]	0.72 (0.67,0.76)
Algorithm + SCORE	0.72 (0.67,0.76)

Of the 12,026 patients in the UK Biobank validation dataset, 91 had experienced a previous cardiac event before retinal imaging and were excluded from the analysis. Of the 11,835 patients in the validation dataset without a previous cardiac event, 105 experienced a MACE within 5 years of retinal imaging. 95% CIs were calculated using 2,000 bootstrap samples.

Discussion

Our results indicate that the application of deep learning to retinal fundus images alone can be used to predict multiple cardiovascular risk factors, including age, gender and SBP. That these risk factors are core components used in multiple cardiovascular risk calculators indicates that our model can potentially predict cardiovascular risk directly. This is supported by our preliminary results for the prediction of MACE, which achieve a similar accuracy to the composite SCORE risk calculator. This is also consistent with previous studies that suggest that retinal imaging contains information about cardiovascular risk factors such as age and blood pressure[35] as well as MACE[18,20,22]. Building on this body of work, we demonstrate not only that these signals are present in the retina, but that they are also quantifiable to a degree of precision not reported before.

Encouragingly, the corresponding attention maps also indicate that the neural-network model is paying attention to the vascular regions in the retina to predict several variables associated with cardiovascular risk. These attention data, together with the fact that our results are consistent in two separate validation datasets, suggest that the predictions are likely to generalize to other datasets, and provide indications of pathological phenomena that can be studied further. For example, our results show strong gender differences in the fundus photographs and may help guide basic research investigating the anatomical or physiological differences between male and female eyes. Similarly, our findings may aid the scientific community in advancing its understanding of how cardiovascular disease processes or risk factors affect the retinal vasculature or optic disc in patients.

Despite the promising results, our study has several limitations. First, it only used images with a 45° field of view, and future work could examine the generalizability of these findings to images with either smaller or larger fields of view. In addition, the overall size of the dataset is relatively small for deep learning. In particular, although the AUC for cardiovascular events was comparable to that of SCORE, the CIs for both methods were wide. A significantly larger dataset or a population with more cardiovascular events may enable more accurate deep-learning models to be trained and evaluated with high confidence. Another limitation is that some important inputs to existing cardiovascular risk calculators were missing from the datasets. In particular, lipid panels were not available for either the UK Biobank or the EyePACS datasets and a 'gold standard' diagnosis of diabetic status was not available in the UK Biobank dataset. Adding such input data is likely to improve the performance of cardiovascular risk prediction for both existing models and those we propose. Similarly, risk factors such as blood pressure and HbA1c were only available in one of the datasets (Table 2). Furthermore, some variables such as smoking status were self-reported and may be biased. In particular, because former smokers were grouped with never-smokers, it is possible

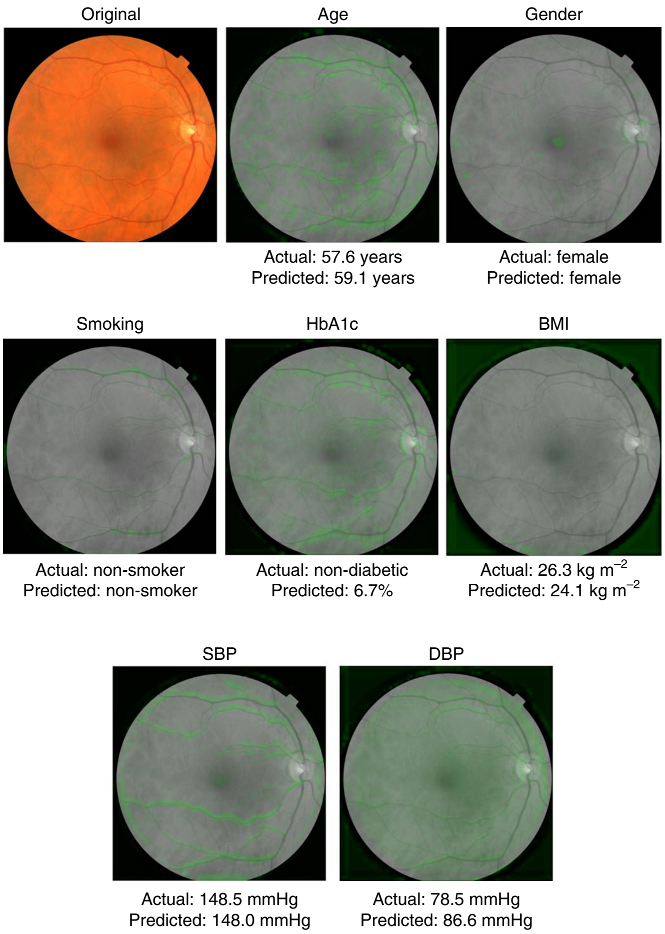

Fig. 2 | Attention maps for a single retinal fundus image. The top left image is a sample retinal image in colour from the UK Biobank dataset. The remaining images show the same retinal image, but in black and white. The soft attention heat map (Methods) for each prediction is overlaid in green, indicating the areas of the heat map that the neural-network model is using to make the prediction for the image. For a quantitative analysis of what was highlighted, see Table 6. HbA1c values are not available for UK Biobank patients, so the self-reported diabetes status is shown instead.

Table 6 | Percentage of the 100 attention heat maps for which doctors agreed that the heat map highlighted the given feature

Risk factor	Vessels (%)	Optic disc (%)	Non-specific features
Age	95	33	38
Gender	71	78	50
Current smoker	91	25	38
HbA1c	78	32	46
SBP	98	14	54
DBP	29	5	97
BMI	1	6	99

Heat maps (n = 100) were generated for each risk factor and then presented to three ophthalmologists who were asked to check the features highlighted in each image (n = 300 responses for each risk factor). The images were shuffled and presented as a set of 700, and the ophthalmologists were blinded to the output prediction of the heat maps and the ground-truth label. For the variables that were present in both datasets (age and gender), the most commonly highlighted features were identical in both datasets.

that the AUC of the model might change if there were different or finer groupings of patients on the basis of length of previous smoking history. Additional validation of our models on other datasets would be beneficial for all these predictions. Training with larger datasets and more clinical validation will also help determine whether retinal fundus images may be able to augment or replace some of the other markers, such as lipid panels, to yield a more accurate cardiovascular risk score.

In summary, we have provided evidence that deep learning may uncover additional signals in retinal images that will allow for better cardiovascular risk stratification. In particular, they could enable cardiovascular assessment at the population level by leveraging the existing infrastructure used to screen for diabetic eye disease. Our work also suggests avenues of future research into the source of these associations, and whether they can be used to better understand and prevent cardiovascular disease.

Methods

Study participants. We used two datasets in this study. Data in the first dataset—the UK Biobank—were generated in an observational study that recruited 500,000 participants, aged 40–69 years, across the UK between 2006 and 2010. Each participant gave consent and went through a series of health measurements and questionnaires. Participants were then followed for health outcomes, such as hospitalizations, mortality and cause of death. The study was reviewed and approved by the North West Multi-Centre Research Ethics Committee.

Detailed protocols for obtaining the measurements from participants are available on the UK Biobank website at www.ukbiobank.ac.uk. Briefly, smoking status was obtained via a survey that was administered using a touchscreen interface. Participants were asked to self-identify as a current smoker, former smoker or never-smoker. Those who had a smoking history were then asked for additional details. For the purpose of this study, the population was binarized into those who were current smokers and those who were not. In addition, measurements of resting blood pressure were obtained with the participant seated using an automated Omron 705-IT electronic blood pressure monitor (Omron Healthcare Europe). Automated blood pressure was obtained twice and the average of the two was used in our study. Each participant also provided blood, urine and saliva samples (http://www.ukbiobank.ac.uk/about-biobank-uk). However, glucose, cholesterol and HbA1c measurements were not available at the time of this study. A total of 67,725 patients (http://www.ukbiobank.ac.uk/eye-vision-consortium/) subsequently underwent paired retinal fundus and optical coherence tomography imaging using a Topcon 3D OCT-1000 MKII (Topcon Corporation). Fundus images from this dataset consisted of images with a 45° primary field of view. We divided the UK Biobank dataset into a development dataset to develop our models (80%) and a validation dataset to assess our models' performance (20%). Images of poor quality were filtered out before training and validation. This accounted for approximately 12% of patients in the UK Biobank dataset.

The second dataset—EyePACS—was generated by a US-based teleretinal-services provider that offers screening for diabetic eye disease to over 300 clinics worldwide. EyePACS images were acquired as part of routine clinical care for diabetic retinopathy screening, and approximately 40% of the images were acquired with pupil dilation. A variety of cameras were used, including CenterVue DRS, Optovue iCam, Canon CR1/DGi/CR2 and Topcon NW using 45° fields of view. In most cases, each eye was imaged three times to capture the primary (roughly macula-centred), nasal (roughly disc-centred) and temporal (temporal to macula) fields. A subset of the EyePACS clinics recorded HbA1c at each visit. All images and data were de-identified according to the Health Insurance Portability and Accountability Act 'Safe Harbor' before transfer to investigators. Ethics review and Institutional Review Board exemption was obtained via the Quorum Review Independent Review Board (Seattle, WA, USA).

Retinal fundus images from the EyePACS dataset collected between 2007 and 2015 were used as our development dataset. For the clinical validation dataset (EyePACS-2K), we used a random sample of macula-centred images taken at EyePACS screening sites between May 2015 and October 2015 with HbA1c measurements (Table 1). There was no overlap in patients between the EyePACS development dataset and the EyePACS-2K validation dataset. This development–validation split ratio was chosen to match the clinical validation dataset used in a previous report[33]. No image-quality filtering was applied to the EyePACS development and clinical validation datasets.

Model development. A deep-neural-network model is a sequence of mathematical operations applied to the input, such as pixel values in an image. There can be millions of parameters (weights) in this mathematical function[36]. Deep learning is the process of learning the correct parameter values (training), such that this function performs a given task. For example, the model can output a prediction of interest from the pixel values in a fundus image. The development dataset is divided into two components: a 'training' dataset and a 'tuning' dataset (the 'tuning' dataset is also commonly called the validation dataset, but we wish to avoid confusion with a clinical validation dataset, which consists of data the model did not train on). During the training process, the parameters of the neural network are initially set to random values. Then, for each image, the prediction given by the model is compared with the known label from the training dataset, and parameters of the model are then modified slightly to decrease the error on that image (stochastic gradient descent). This process is repeated for every image in the training dataset until the model 'learns' how to accurately compute the label from the pixel intensities of the image for all images in the training dataset. With appropriate tuning and sufficient data, the result is a model general enough to predict the labels (for example, cardiovascular risk factors) on new images. In this study, we used the Inception-v3 neural-network architecture[37] to predict the labels.

We pre-processed the images for training and validation and trained the neural network following a previously reported procedure[31], but for multiple predictions simultaneously: age, gender, smoking status, BMI, SBP, DBP and HbA1c. Input images were scale-normalized by detecting the circular mask of the fundus image and resizing the diameter of the fundus to be 587 pixels wide. Images for which the circular mask could not be detected or those that were of poor quality were excluded from training. The optimization algorithm used to train the network weights was a distributed-stochastic-gradient-descent implementation[38]. To speed up the training, batch normalization[39], as well as pre-initialization using weights from the same network trained to classify objects in the ImageNet dataset[40], were used.

To keep the loss functions on consistent scales, we trained three separate models—a 'classification model' for predicting the binary risk factors (gender and smoking status), a 'regression model' for the continuous risk factors (age, BMI, blood pressures and HbA1c) and a third classification network for predicting MACE. The MACE model was separated from the classification model because the scarcity of the outcome prompted us to augment the binary variables with discretized (binned) versions of the continuous variables as additional prediction heads. Specifically, we used the following cut-offs: <120, 120–140, 140–160 and ≥ 160 for SBP; <50, 50–60 and ≥60 for age; and <18.5, 18.5–25, 25–30, 30–35, 35–40 and ≥40 for BMI. To optimize for generalizability, we pooled the UK Biobank and EyePACS datasets for model development. The MACE model was only trained using the UK Biobank dataset because MACE outcomes were only available for that dataset.

Because the network in this study had a large number of parameters (22 million), we used early stopping criteria[41] to help avoid overfitting: the termination of training when the model performance (such as AUC; see 'Statistical analysis' in Methods) on a 'tuning dataset' stopped improving. The tuning dataset was a random subset of the development dataset that was used as a small evaluation dataset for tuning the model rather than to train the model parameters. This tuning dataset comprised 10% of the UK Biobank dataset and 2.1% of the EyePACS dataset. To further improve the results, we averaged the results of ten neural-network models that were trained on the same data (ensembling[42]).

TensorFlow (http://tensorflow.org), an open-source software library for machine intelligence, was used in the training and evaluation of the models.

Evaluating the algorithm. To evaluate the model performance for continuous predictions (age, SBP, DBP and HbA1c), we used the MAE and the coefficient of determination (R^2). For binary classification (gender and smoking status), we used the AUC. For multiclass classification, we used a simple Cohen's κ. Images in the clinical validation dataset were all 45° fundus photographs with a primary field of view. Images of poor quality were excluded from the clinical validation datasets.

Statistical analysis. To assess the statistical significance of the results, we used the non-parametric bootstrap procedure: from the validation dataset of n patients, we sampled n patients with replacement and evaluated the model on this sample. By repeating this sampling and evaluation 2,000 times, we obtained a distribution of the performance metric (such as AUC) and reported the 2.5 and 97.5 percentiles as 95% CIs.

To further assess the statistical significance of the performance of the models for predicting continuous risk factors such as age and SBP, we used a one-tailed binomial test for the 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 (for example, size 10 for ±5) across the population histogram and then taking the maximum of the summed histogram counts. This provides the maximum possible 'random' accuracy (by guessing the centre of the sliding window containing the maximum probability mass).

Mapping attention. To better understand how the neural-network models arrived at the predictions, we used a deep-learning technique called soft attention[43–45]. Briefly, we used the following architecture: the input images were 587 × 587 pixels and the saliency map was originally 73 × 73 pixels. There were 3 2 × 2 maxpool layers and 4 3 × 3 convolutional layers before the saliency map, as well as 3 reverse maxpool layers to upscale the saliency map from 73 × 73 pixels back to 587 × 587 pixels. The convolutional layers contained 64, 128, 256 and 512 filters. The path from input image to saliency map is described in Supplementary Table 1. The first two dimensions are image size. The third is number of filters (or channels in the case of the input image).

This technique is described in more detail elsewhere[42]. These small models are less powerful than Inception-v3. They were used only for generating attention heat maps and not for the best performance results observed with Inception-v3. For each prediction shown in Fig. 2, a separate model with identical architecture was trained. The models were trained on the same training data as the Inception-v3 network described above, and the same early stopping criteria were used.

To provide a qualitative assessment of the features that are highlighted in the heat maps, we generated 100 images for each of the predicted factors from 3 image sets for a total of 700 images. For the BMI, current smoker, SBP and DBP predictions, we randomly sampled 100 images for each of these predictions from the UK Biobank dataset. For HbA1c, we randomly sampled 100 images from the EyePACS dataset. For age and gender, we randomly sampled 50 images from the EyePACS dataset and 50 from the UK Biobank dataset. The 700 images were shown to three ophthalmologists in the same (randomized) order using a survey form (see Supplementary Fig. 1 for a screenshot of the form) for a total of 300 responses per prediction. On the basis of feedback from the ophthalmologists, we aggregated their responses so that veins, arteries, arterioles, venules and vessel surroundings were reported as 'vessels', optic disc and optic-disc edges were reported as 'optic disc' and image edges and 'nothing in particular' were reported as 'non-specific features'. The macula was not one of the checkbox options, but ophthalmologists repeatedly reported highlighting of the macula for the gender predictions.

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Code availability. The machine-learning models were developed and deployed using standard model libraries and scripts in TensorFlow. Custom code was specific to our computing infrastructure and mainly used for data input/output and parallelization across computers.

Data availability. The data that support the findings of this study are available, with restrictions, from UK Biobank and EyePAC.

Received: 23 August 2017; Accepted: 12 January 2018;
Published online: 19 February 2018

References

1. WHO *The Top 10 Causes of Death* (2017).
2. Stone, N. J. et al. 2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation* **129**, S1–S45 (2014).
3. Wilson, P. W. et al. Prediction of coronary heart disease using risk factor categories. *Circulation* **97**, 1837–1847 (1998).
4. National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. *Circulation* **106**, 3143–3421 (2002).
5. D'Agostino, R. B. Sr et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation* **117**, 743–753 (2008).
6. Conroy, R. M. et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *Eur. Heart J.* **24**, 987–1003 (2003).
7. Graham, I. et al. Fourth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of nine societies and by invited experts). *Eur. J. Cardiovasc. Prev. Rehabil.* **14**, E1–E40 (2007).
8. Yeboah, J. et al. Comparison of novel risk markers for improvement in cardiovascular risk assessment in intermediate-risk individuals. *JAMA* **308**, 788–795 (2012).
9. Goff, D. C. et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation* **129**, S49–S73 (2014).
10. Hira, R. S. et al. Frequency and practice-level variation in inappropriate aspirin use for the primary prevention of cardiovascular disease: insights from the National Cardiovascular Disease Registry's Practice Innovation and Clinical Excellence registry. *J. Am. Coll. Cardiol.* **65**, 111–121 (2015).
11. *Cardiovascular Disease (10-Year Risk)* (Framingham Heart Study, accessed 21 June 2017); https://www.framinghamheartstudy.org/risk-functions/cardiovascular-disease/10-year-risk.php
12. Cooney, M. T. et al. How much does HDL cholesterol add to risk estimation? A report from the SCORE investigators. *Eur. J. Cardiovasc. Prev. Rehabil.* **16**, 304–314 (2009).
13. Dudina, A. et al. Relationships between body mass index, cardiovascular mortality, and risk factors: a report from the SCORE investigators. *Eur. J. Cardiovasc. Prev. Rehabil.* **18**, 731–742 (2011).

14. Wang, J. J. et al. Retinal vascular calibre and the risk of coronary heart disease-related death. *Heart* **92**, 1583–1587 (2006).
15. Wong, T. Y. et al. Quantitative retinal venular caliber and risk of cardiovascular disease in older persons: the cardiovascular health study. *Arch. Intern. Med.* **166**, 2388–2394 (2006).
16. Seidelmann, S. B. et al. Retinal vessel calibers in predicting long-term cardiovascular outcomes: the Atherosclerosis Risk in Communities Study. *Circulation* **134**, 1328–1338 (2016).
17. Wong, T. Y. et al. Retinal vascular caliber, cardiovascular risk factors, and inflammation: the multi-ethnic study of atherosclerosis (MESA). *Invest. Ophthalmol. Vis. Sci.* **47**, 2341–2350 (2006).
18. McGeechan, K. et al. Meta-analysis: retinal vessel caliber and risk for coronary heart disease. *Ann. Intern. Med.* **151**, 404–413 (2009).
19. McGeechan, K. et al. Prediction of incident stroke events based on retinal vessel caliber: a systematic review and individual-participant meta-analysis. *Am. J. Epidemiol.* **170**, 1323–1332 (2009).
20. Wong, T. Y. et al. Retinal arteriolar narrowing and risk of coronary heart disease in men and women. The Atherosclerosis Risk in Communities Study. *JAMA* **287**, 1153–1159 (2002).
21. Witt, N. et al. Abnormalities of retinal microvascular structure and risk of mortality from ischemic heart disease and stroke. *Hypertension* **47**, 975–981 (2006).
22. Wong, T. Y. et al. Retinal microvascular abnormalities and 10-year cardiovascular mortality: a population-based case–control study. *Ophthalmology* **110**, 933–940 (2003).
23. Cheung, C. Y.-L. et al. Retinal microvascular changes and risk of stroke: the Singapore Malay Eye Study. *Stroke* **44**, 2402–2408 (2013).
24. Liew, G. et al. Fractal analysis of retinal microvasculature and coronary heart disease mortality. *Eur. Heart J.* **32**, 422–429 (2011).
25. Kawasaki, R. et al. Fractal dimension of the retinal vasculature and risk of stroke: a nested case–control study. *Neurology* **76**, 1766–1767 (2011).
26. Cheung, C. Y. et al. Retinal vascular fractal dimension and its relationship with cardiovascular and ocular risk factors. *Am. J. Ophthalmol.* **154**, 663–674.e1 (2012).
27. Mookiah, M. R. K. et al. Computer-aided diagnosis of diabetic retinopathy: a review. *Comput. Biol. Med.* **43**, 2136–2155 (2013).
28. Roychowdhury, S., Koozekanani, D. D. & Parhi, K. K. DREAM: diabetic retinopathy analysis using machine learning. *IEEE J. Biomed. Health Inform.* **18**, 1717–1728 (2014).
29. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
30. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
31. 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).
32. Ting, D. S. W. et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA* **318**, 2211–2223 (2017).
33. Krause, J. et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. Preprint at https://arxiv.org/abs/1710.01711 (2017).
34. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: visualising image classification models and saliency maps. Preprint at https://arxiv.org/abs/1312.6034 (2013).
35. Wong, T. Y., Klein, R., Klein, B. E. K., Meuer, S. M. & Hubbard, L. D. Retinal vessel diameters and their associations with age and blood pressure. *Invest. Ophthalmol. Vis. Sci.* **44**, 4644–4650 (2003).
36. Angermueller, C., Pärnamaa, T., Parts, L. & Stegle, O. Deep learning for computational biology. *Mol. Syst. Biol.* **12**, 878 (2016).
37. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. Preprint at https://arxiv.org/abs/1512.00567 (2015).
38. Dean, J. et al. Large scale distributed deep networks. In *Proc. 25th Conference on Advances in Neural Information Processing Systems* (eds Pereira, F. et al.) 1223–1231 (Neural Information Processing Systems Foundation, 2012).
39. Ioffe, S. & Szegedy, C. Batch normalization: accelerating deep network training by reducing internal covariate shift. Preprint at https://arxiv.org/abs/1502.03167 (2015).
40. Russakovsky, O. et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).
41. Caruana, R., Lawrence, S. & Giles, L. Overfitting in neural nets: backpropagation, conjugate gradient, and early stopping. In *Proc. 13th Conference on Advances in Neural Information Processing Systems* (eds Leen, T. K. et al.) (MIT Press, Cambridge, MA, 2001).
42. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *Proc. 25th Conference on Advances in Neural Information Processing Systems* (eds Pereira, F. et al.) 1097–1105 (Neural Information Processing Systems, 2012).
43. Xu, K. et al. Show, attend and tell: neural image caption generation with visual attention. Preprint at https://arxiv.org/abs/1502.03044 (2015).
44. Cho, K., Courville, A. & Bengio, Y. Describing multimedia content using attention-based encoder–decoder networks. *IEEE Trans. Multimed.* **17**, 1875–1886 (2015).
45. Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. Preprint at https://arxiv.org/abs/1409.0473 (2014).

Acknowledgements

We thank C. Angermueller, A. Narayanaswamy, A. Rajkomar, A. Taly and P. Nelson from Google Research for their advice and assistance with reviewing the manuscript, and J. Cuadros for access to and clarifications on the EyePACS dataset and for thoughtful discussions. We used the UK Biobank Resource under application number 17643 for this work.

Author contributions

R.P., A.V.V., Y.L., G.S.C., L.P. and D.R.W. designed the research; R.P., A.V.V., K.B., Y.L. and L.P. acquired data and/or executed the research; R.P., A.V.V., K.B., Y.L., M.V.M., L.P. and D.R.W. analysed and/or interpreted the data; R.P., A.V.V., K.B., Y.L., M.V.M., G.S.C., L.P. and D.R.W. prepared the manuscript.

Competing interests

The authors are employees of Google and Verily Life Sciences.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41551-018-0195-0.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to L.P.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Lily Peng

□ Initial submission □ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

On the basis of our previous experience and published literature, we know that deep learning requires on the order of tens of thousands or hundreds of thousands of examples. As such, we included as much available data as possible from these datasets.

2. Data exclusions

Describe any data exclusions.

For the validation set, we excluded any images that were of poor quality or with missing data. These are pre-established exclusions.

3. Replication

Describe whether the experimental findings were reliably reproduced.

We ensured that our results generalized over two distinct datasets that were not used for training.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Samples were randomly allocated to training, test, and validation datasets.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

This is a retrospective study. Splits for validation were random and automatically generated. No blinding was necessary.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
□	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	A statement indicating how many times each experiment was replicated
□	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☒	□	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
□	☒	The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
□	☒	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☒	□	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	TensorFlow and python scripts.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	No unique materials were used.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	No cell lines were used.
b. Describe the method of cell line authentication used.	No cell lines were used.
c. Report whether the cell lines were tested for mycoplasma contamination.	No cell lines were used.
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	No animals were used.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	Retina fundus images were obtained from an adult diabetic-screening population in the US (EyePACS) and from the general population in the UK.