Predicting the risk of developing diabetic retinopathy using deep learning

Ashish Bora, Siva Balasubramanian, Boris Babenko, Sunny Virmani, Subhashini Venugopalan, Akinori Mitani, Guilherme de Oliveira Marinho, Jorge Cuadros, Paisan Ruamviboonsuk, Greg S Corrado, Lily Peng, Dale R Webster, Avinash V Varadarajan, Naama Hammel, Yun Liu, Pinal Bavishi**

Summary

Background Diabetic retinopathy screening is instrumental to preventing blindness, but scaling up screening is challenging because of the increasing number of patients with all forms of diabetes. We aimed to create a deep-learning system to predict the risk of patients with diabetes developing diabetic retinopathy within 2 years.

Methods We created and validated two versions of a deep-learning system to predict the development of diabetic retinopathy in patients with diabetes who had had teleretinal diabetic retinopathy screening in a primary care setting. The input for the two versions was either a set of three-field or one-field colour fundus photographs. Of the 575 431 eyes in the development set 28 899 had known outcomes, with the remaining 546 532 eyes used to augment the training process via multitask learning. Validation was done on one eye (selected at random) per patient from two datasets: an internal validation (from EyePACS, a teleretinal screening service in the USA) set of 3678 eyes with known outcomes and an external validation (from Thailand) set of 2345 eyes with known outcomes.

Findings The three-field deep-learning system had an area under the receiver operating characteristic curve (AUC) of 0·79 (95% CI 0·77–0·81) in the internal validation set. Assessment of the external validation set—which contained only one-field colour fundus photographs—with the one-field deep-learning system gave an AUC of 0·70 (0·67–0·74). In the internal validation set, the AUC of available risk factors was 0·72 (0·68–0·76), which improved to 0·81 (0·77–0·84) after combining the deep-learning system with these risk factors (p<0·0001). In the external validation set, the corresponding AUC improved from 0·62 (0·58–0·66) to 0·71 (0·68–0·75; p<0·0001) following the addition of the deep-learning system to available risk factors.

Interpretation The deep-learning systems predicted diabetic retinopathy development using colour fundus photographs, and the systems were independent of and more informative than available risk factors. Such a risk stratification tool might help to optimise screening intervals to reduce costs while improving vision-related outcomes.

Funding Google.

Lancet Digit Health 2021; 3: e10–19

Published Online November 26, 2020 https://doi.org/10.1016/S2589-7500(20)30250-8

*Contributed equally

Google Health (A Bora MS, B Babenko PhD, S Virmani MS, A Mitani MD, G de Oliveira Marinho BS, G S Corrado PhD, L Peng MD, D R Webster PhD, A V Varadarajan MS, N Hammel MD, Y Liu PhD, P Bavishi BE) **and Google Research** (S Venugopalan PhD), **Google, Mountain View, CA, USA; Advanced Clinical, Deerfield, IL, USA** (S Balasubramanian MD); **EyePACS, Santa Cruz, CA, USA** (J Cuadros OD); **and Department of Ophthalmology, Rajavithi Hospital, College of Medicine, Rangsit University, Bangkok, Thailand** (Prof P Ruamviboonsuk MD)

Correspondence to: Dr Naama Hammel, Google Health, Google, Mountain View, CA 94043, USA **nhammel@google.com**

Introduction

Globally, diabetic retinopathy is the leading cause of preventable blindness in adults aged 20–74 years.[1] With the goal of early detection, regular screening is recommended by major organisations—including the American Diabetes Association,[2] International Council of Ophthalmology,[3] and American Academy of Ophthalmology[4]—at intervals ranging from every 12 months to 24 months for patients with no or mild diabetic retinopathy. Although regular screening is crucial to preventing blindness, the expected increase in the number of patients with diabetes—from 415 million in 2015 to a predicted 642 million in 2040[5]—means that the burden of screening and follow-up represent a substantial challenge. The efficiency of diabetic retinopathy screening programmes might be improved by personalising screening frequencies on the basis of the likelihood of the development or progression of diabetic retinopathy.[6] We created a deep-learning system that uses colour fundus photographs to predict the risk of developing diabetic retinopathy.

Signs of retinal microvascular changes caused by diabetes are visible in colour fundus photographs, which are routinely used to assess the stage of diabetic retinopathy. In addition to disease stage, the risk of development and progression of diabetic retinopathy is influenced by several risk factors. Modifiable risk factors include hyperglycaemia, hypertension, dyslipidaemia and obesity, smoking, anaemia, pregnancy, low health literacy, inadequate access to health care, and poor adherence to therapy.[7,8] Non-modifiable risk factors include ethnicity, family history or genetics, age at onset of diabetes, type of diabetes, and duration of diabetes.[7]

We used both colour fundus photographs and known risk factors to improve risk stratification for developing diabetic retinopathy. This is a challenging task because microvascular changes are not known to be detectable in colour fundus photographs before the development of diabetic retinopathy. Consequently, grading systems based on colour fundus photographs—such as the Early Treatment Diabetic Retinopathy Study,[9] International

Research in context

Evidence before this study

We searched Google Scholar, PubMed, Scopus, MEDLINE, and Web of Science from the inception of each database until June 15, 2020, for studies published in English using the search terms "progression of diabetic retinopathy", "predicting development of diabetic retinopathy", "predicting incident diabetic retinopathy", "risk factors for diabetic retinopathy", "diabetic retinopathy screening programs", "automated detection of diabetic retinopathy using fundus photographs", "artificial intelligence", "machine learning", and "deep learning". Although few studies have investigated the risk of diabetic retinopathy progression on fundus photographs using deep learning, to our knowledge no study has reported automated risk stratification of developing diabetic retinopathy from no diabetic retinopathy on fundus photographs or this prediction in combination with known risk factors.

Added value of this study

In this study, we created a deep-learning system that used colour fundus photographs to predict the development of mild or worse diabetic retinopathy within 2 years, for patients with diabetes who did not have diabetic retinopathy at the time of screening. The deep-learning system was created using a large retrospective longitudinal dataset collected in a diabetic retinopathy screening programme. Significant improvement in risk stratification was observed in two independent datasets from two different countries. The use of an automated risk stratification tool could help to scale diabetic retinopathy screening as the number of patients with diabetes continues to grow.

Implications of all the available evidence

Our risk stratification tool could be used by clinicians to optimise diabetic retinopathy screening intervals. Patients with diabetes identified by our algorithm as being at high risk could be followed up more closely to improve visual outcomes, with patients at lower risk followed up less frequently to reduce the burden of screening. More effective screening strategies are particularly important given the challenges introduced by the COVID-19 pandemic. The risk prediction, together with personalised patient education could motivate patients to undertake lifestyle modifications or increase their adherence to therapy for improved blood sugar control. Finally, our tool could aid research into diabetes management and treatment. Prospective studies are required to validate the deep-learning system as a risk predictor for diabetic retinopathy development and its effectiveness in real-world screening programmes and clinical settings.

Clinical Diabetic Retinopathy,[10] and the Royal College of Ophthalmologists Diabetic Retinopathy guidelines[11]—do not stratify patients who do not have diabetic retinopathy. The subgroup of patients without diabetic retinopathy is important because most patients who are screened for diabetic retinopathy show no signs of the disease.[12] Furthermore, only a small proportion of these patients require retinal intervention within 2 years,[13] highlighting the need to identify patients at high risk of developing diabetic retinopathy and of disease progression. We hypothesised that a deep-learning system could provide such stratification by evaluating colour fundus photographs of the eyes of patients without diabetic retinopathy and predicting whether the patient would develop mild or worse disease within 2 years.

Methods

Study population and datasets

We used deidentified colour fundus photographs from two longitudinal datasets to develop and evaluate the deep-learning system. The first dataset included 362 283 patients who attended a total of 430 917 visits at 759 sites from EyePACS, a teleretinal diabetic retinopathy screening service in the USA. These images were acquired during routine clinical care for diabetic retinopathy screening. The patients from the EyePACS dataset were randomly split (8:2) in two, with no overlap between the development dataset and the internal validation set (72 457 [20%]). The deep-learning system was validated using the internal validation set, which contained 9124 eyes after selecting one eye at random per patient using a pseudorandom number generator and ensuring that there were at least two sets of gradable images on different visits. 7976 eyes remained after filtering for those without diabetic retinopathy at the first visit.

See Online for appendix

The second dataset was from 13 medical centres from across Thailand and included 6791 patients.[14] Ruamviboonsuk and colleagues[14] randomly selected the 6791 patients from the national diabetic patients registry in Thailand and represented hospitals and health centres from each of the 13 health regions in Thailand. The standard imaging protocol involved taking 45° primary field colour fundus photographs (appendix p 1). Most patients were followed up after approximately 2 years. One eye (selected at random using a pseudorandom number generator) per patient was assessed and both visits for the eye were graded (appendix p 2). When considering only patients with two gradable visits and no diabetic retinopathy at the first visit, 4762 eyes were included in the external validation set.

All images and metadata were deidentified according to the Health Insurance Portability and Accountability Act[15] Safe Harbor provision before transfer to the study investigators. Ethics review and institutional review

board exemption was obtained from Advarra Institutional Review Board.

Outcomes

The primary outcome was to predict the development of diabetic retinopathy, defined as whether an eye without diabetic retinopathy (according to grading scale; appendix p 2) at baseline was graded as having mild or worse diabetic retinopathy within 2 years, with a 6 month buffer to account for visits close to this date (appendix pp 1–2).Eyes without a known 2-year mild or worse diabetic retinopathy endpoint were considered for the survival analysis.

Algorithm development

The deep-learning system was developed based on the Inception-v3 architecture[16,17] and trained on the training set (which contained 88% of the patients included in the development dataset) and tuned using the tuning set (which contained 12% of the images included in the development set). We developed two versions of the deep-learning system, one that took the primary field as input (one-field) and a second that took nasal, primary, and temporal images as input (three-field). The three-field deep-learning system processed each field using an identical Inception-v3 module (ie, with shared weights), and the output feature vectors are concatenated before being used as input to the final classification layer. We present the results for both versions of the deep-learning system for the internal validation set. However, the external validation set contained only the primary field, and thus only allowed evaluation of the one-field deep-learning system. Both versions of the deep-learning

	Development dataset		Validation datasets	
	Training dataset	Tuning dataset	Internal validation dataset (EyePACS)	External validation dataset (Thailand)
Patients	253 598	36 228	72 457	6791
Total number of eyes	503 527	71 904	143 864	13 573
Eyes with at least one set of images with available grades	365 144 (72·5%)	52 456 (73·0%)	104 695 (72·8%)	6307 (46·5%)
Eyes after selecting random eye per patient for validation purposes	365 144 (72·5%)	52 456 (73·0%)	52 403 (36·4%)	6 307 (46·5%)
Eyes with at least two sets of gradable images (on different visits)	63 281 (12·6%)	9333 (13·0%)	9124 (6·3%)	5603 (41·3%)
Eyes with at least two sets of gradable images and without diabetic retinopathy at the first visit (%; number of patients)	55 561 (11·0%; 31 449)	8211 (11·4%; 4651)	7976 (5·5%; 7976)	4762 (35·1%; 4762)
Women* (%)	19 919/31 447 (63·3%)	2905/4649 (62·5%)	5118/7975 (64·2%)	3277/4759 (68·9%)
Men* (%)	11 528/31 447 (36·7%)	1744/4649 (37·5%)	2857/7975 (35·8%)	1482/4759 (31·1%)
Median age, years (IQR)	53 (46–60)	54 (45–60)	54 (45–60)	59 (52–66)
Race and ethnicity (%)				
Hispanic	32 779 (59·0%)	4790 (58·3%)	4681 (58·7%)	0
White	4077 (7·3%)	605 (7·4%)	606 (7·6%)	0
Asian or Pacific Islander	2935 (5·3%)	403 (4·9%)	381 (4·8%)	4762 (100·0%)
Black	2242 (4·0%)	310 (3·8%)	369 (4·6%)	0
Native American	718 (1·3%)	100 (1·2%)	92 (1·2%)	0
Other	861 (1·5%)	135 (1·6%)	136 (1·7%)	0
Not available	11 949 (21·5%)	1868 (22·7%)	1711 (21·5%)	0
Glycated haemoglobin (%), median (IQR)	6·5% (7·3–8·9)	6·5% (7·2–8·9)	6·5% (7·3–9·0)	6·4% (7·2–8·4)
Mild or worse diabetic retinopathy within 2 years				
Eyes with known outcome	25 211	3688	3678	2345
Eyes with mild or worse outcome (%)	4529 (18·0%)	635 (17·2%)	685 (18·6%)	346 (14·8%)
Moderate or worse diabetic retinopathy within 2 years				
Eyes with known outcome	24 393	3609	3554	2292
Eyes with moderate or worse outcome (%)	2093 (8·6%)	303 (8·4%)	317 (8·9%)	225 (9·8%)
Vision-threatening diabetic retinopathy† within 2 years				
Eyes with known outcome	24 104	3602	3500	2219
Eyes with vision-threatening diabetic retinopathy (%)	256 (1·1%)	37 (1·0%)	43 (1·2%)	69 (3·1%)

*A small number of patients did not have this information available. †Includes diabetic macular oedema.

Table 1: **Baseline characteristics**

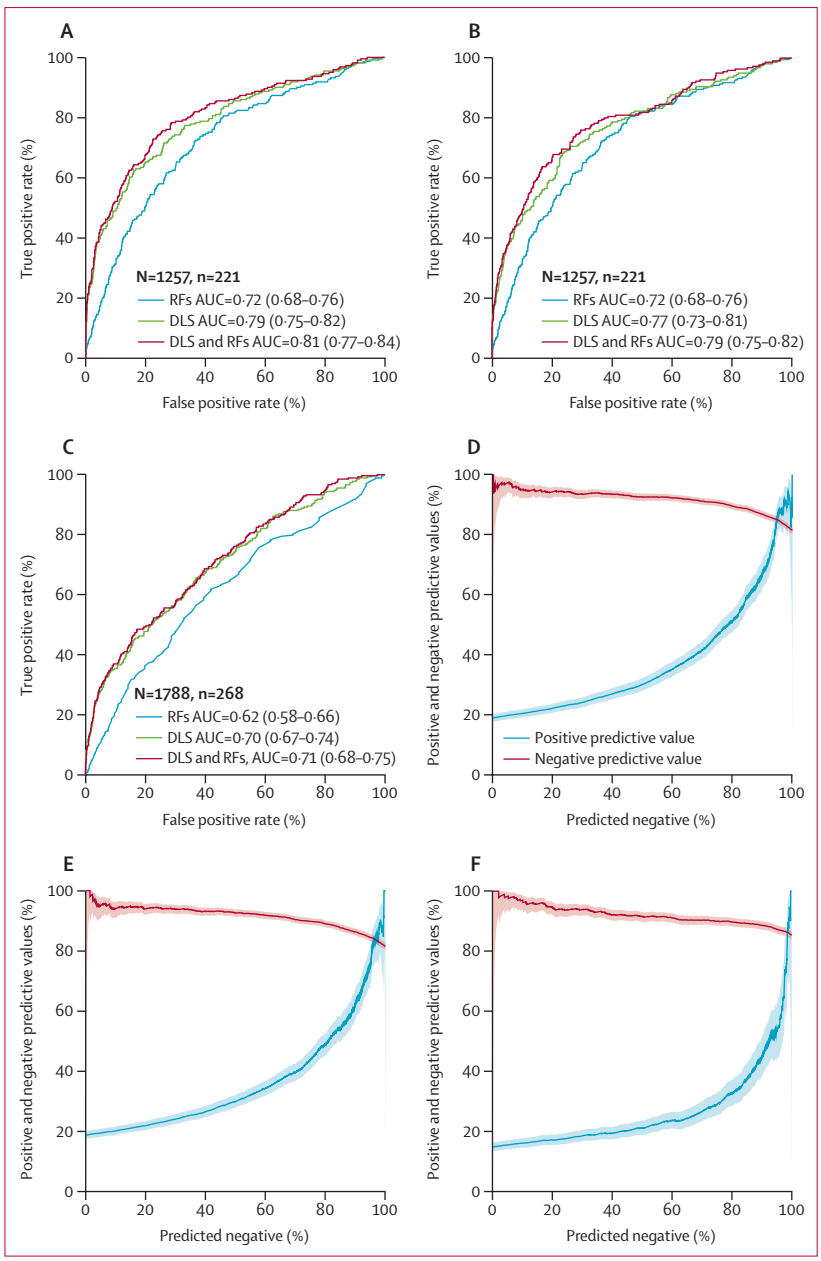


Figure 1: **Discrimination of the DLS for predicting incidence of diabetic retinopathy**
Receiver operating characteristic curves for the three-field (A) and one-field (B) DLS using the internal validation set. (C) Receiver operating characteristic curve for the one-field DLS for the external validation set. The improvements in AUC after adding the DLS to known RFs are all significant (p<0·0001). Curves are for the DLS alone, RFs alone, and for the combination of RFs and the DLS. Receiver operating characteristic curves were plotted only for patients with known values for the RFs (plot for the entire validation set is in the appendix p 15). (D) Positive predictive value and negative predictive value plots for the three-field DLS assessment using the internal validation set. (E) Positive predictive value and negative predictive value plots for the one-field DLS assessment using the internal validation set. (F) Positive predictive value and negative predictive value plots for the one-field DLS assessment using the external validation set. The shaded regions indicate 95% CIs. AUC=area under the curve. DLS=deep-learning system. RFs=risk factors.

system took 587×587 pixel colour fundus photographs as the input and output a number between 0 and 1 representing the likelihood of an eye developing diabetic retinopathy within 2 years (appendix p 2).

Statistical analysis

We evaluated the deep-learning system's ability to predict the development of diabetic retinopathy in multiple ways. We evaluated the discriminatory ability using the area under the receiver operating characteristic curve (AUC) and the system's calibration by plotting the observed event rate against the predicted event rate, based on deciles of predicted risk. Additionally, we evaluated the positive and negative predictive values of the system's predictions at all percentiles of the prediction. Confidence intervals for sensitivity, specificity, and positive predictive values and negative predictive values were calculated using the Clopper-Pearson method based on the β distribution (statsmodels library); confidence intervals for the AUC were calculated with the DeLong method. Statistical significance was evaluated using the DeLong method for the AUC.

To evaluate the deep-learning system with respect to outcomes that occurred after 2 years, we did a survival analysis with Kaplan-Meier curves, log-rank tests, and Cox proportional hazards regression models. For the Kaplan-Meier curves and log-rank tests, thresholds for being at high risk or low risk of developing diabetic retinopathy were based on the upper and lower quartiles of deep-learning system prediction in the tuning set. Survival analysis was done using the lifelines library. Statistical significance was evaluated using the log-rank test for Kaplan-Meier analyses and the likelihood ratio test for Cox analyses.

To compare prognostication by the deep-learning system to that by risk factors, we trained univariable and multivariable logistic regression models on the development set and evaluated them on the validation sets. To ensure that the deep-learning system was not overly confident on the development set, we evaluated the calibration (appendix p 20). Different risk factors were available in each dataset; each experiment included only the patients who had available data for those risk factor(s). For the internal validation set, the risk factors with available data were glycated haemoglobin, self-reported diabetic control, years with diabetes, and insulin use. Diabetic control was reported as poor, fair, moderate, good, or excellent. Insulin use was defined as either patient self-reported or extracted from the electronic medical record when present. For the external validation set, the risk factors with available information were glycated haemoglobin and hypertension status. Categorical variables were encoded as dummy variables. The risk factors were also processed to handle extreme outlier values (appendix p 1). Development and analysis of the saliency maps are detailed in the appendix (p 4).

All analyses were done between April 19, 2019, and June 30, 2020, using Python (version 3.6.7) with the following libraries: numpy (version 1.16.4); pandas (version 0.24.2); seaborn (version 0.9.0); sklearn (version 0.21.3); matplotlib (version 3.0.3); lifelines (version 0.24.6); statsmodels (version 0.10.1); and scipy (version 1.2.1).

	Eyes with risk factor data available	Eyes with diabetic retinopathy within 2 years	Risk factor(s) alone AUC (95%CI)	One-field CFP AUC (95% CI)		Three-field CFP AUC (95% CI)	
				DLS alone	Risk factor(s) and DLS	DLS alone	Risk factor(s) and DLS
Internal validation dataset							
Glycated haemoglobin	2534	458 (18%)	0·68 (0·65–0·71)	0·78 (0·75–0·80)	0·79 (0·77–0·82)	0·79 (0·76–0·81)	0·81 (0·78–0·83)
Years with diabetes	3223	615 (19%)	0·64 (0·61–0·66)	0·77 (0·75–0·79)	0·78 (0·76–0·80)	0·79 (0·76–0·81)	0·80 (0·77–0·82)
Self-reported diabetic control	1589	274 (17%)	0·64 (0·61–0·68)	0·77 (0·73–0·80)	0·78 (0·75–0·81)	0·79 (0·76–0·82)	0·80 (0·77–0·84)
Insulin use	3678	685 (19%)	0·59 (0·58–0·61)	0·78 (0·76–0·80)	0·78 (0·76–0·80)	0·79 (0·77–0·81)	0·80 (0·78–0·82)
Glycated haemoglobin, years with diabetes, self-reported diabetic control, insulin use	1257	221 (18%)	0·72 (0·68–0·76)	0·77 (0·73–0·81)	0·79 (0·75–0·82)	0·79 (0·75–0·82)	0·81 (0·77–0·84)
External validation dataset							
Glycated haemoglobin	1788	268 (15%)	0·62 (0·58–0·66)	0·70 (0·67–0·74)	0·71 (0·68–0·75)	..	..

These risk factors could be used by physicians for risk stratification and represent a useful baseline for evaluating the added value of the DLS. Missing data were not imputed; only patients with available risk factor(s) were included in the analysis in each row. A more comprehensive set of risk factors is reported in appendix (p 10). The only risk factor in the external validation set that was also available and prognostic in the development set was glycated haemoglobin. AUC=area under the curve. CFP=colour fundus photograph. DLS=deep-learning system.

Table 2: **Predictive performance by the DLS applied to CFPs, in comparison with and in combination with known risk factors**

Role of the funding source

Employees of Meta designed and did the study; managed, analysed, and interpreted the data; prepared, reviewed, and approved the Article; and were involved in the decision to submit the Article. All authors affiliated with Google had access to the raw data, and AB and PB verified the data. JC and PR had access to their respective institution's data. The corresponding author had final responsibility for the decision to submit for publication.

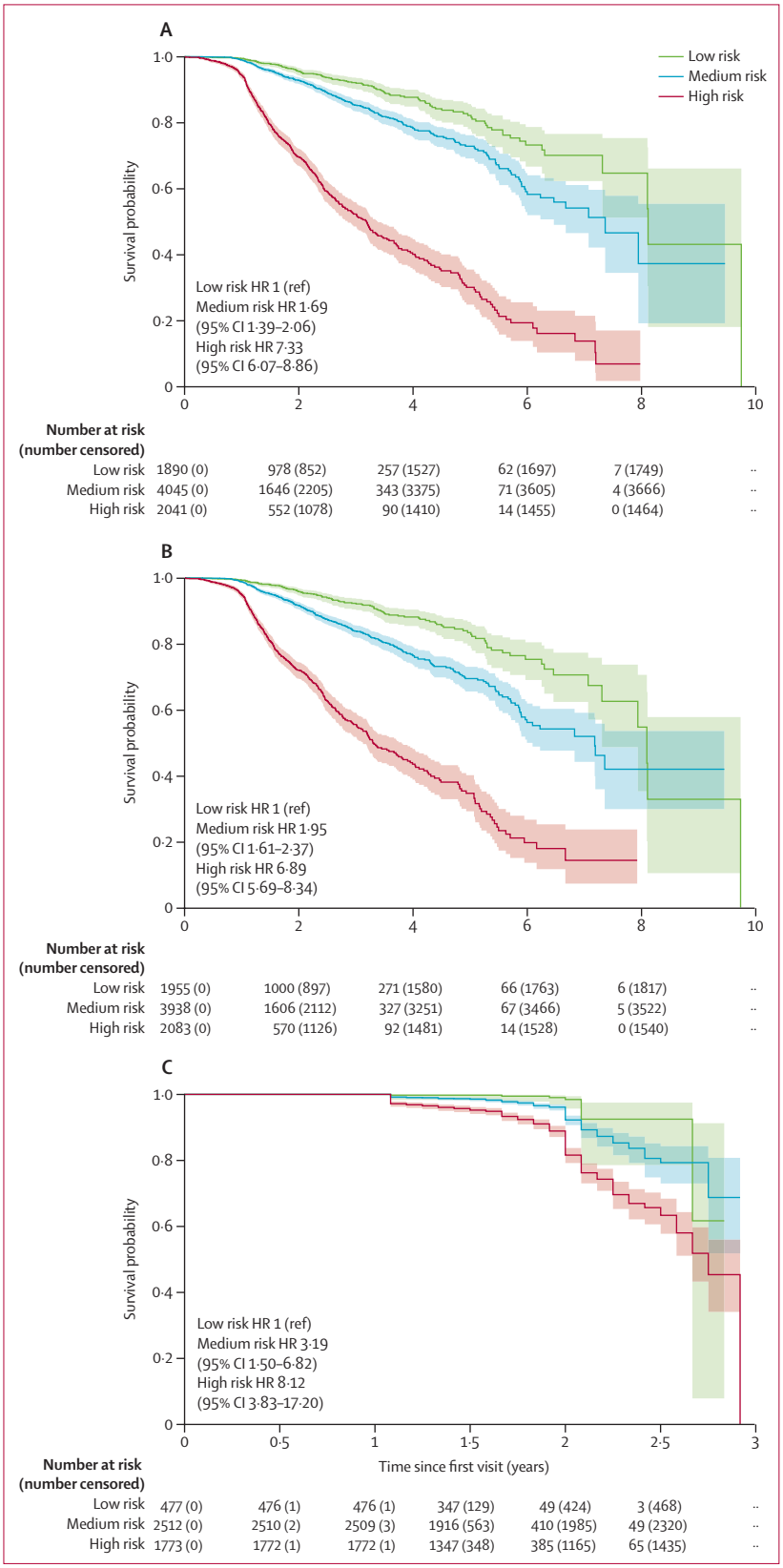
Results

We developed the deep-learning system using the development set that included a total of 575 431 eyes. The system was then evaluated using an internal validation set of 7976 eyes and an external validation set of 4762 eyes. The baseline characteristics, demographics, and distribution of the diabetic retinopathy grades of the development and validation sets are shown in table 1. 685 (19%) of 3678 eyes in the internal validation set and 346 (15%) of 2345 eyes in the external validation set with known 2-year mild or worse diabetic retinopathy outcomes did not have diabetic retinopathy when screened but developed the disease within 2 years. In addition, the two datasets also differed in terms of race and ethnicity (table 1), glycated haemoglobin concentrations (appendix p 14), and grading protocols (appendix pp 6–7).

The three-field AUC for predicting the development of diabetic retinopathy in the internal validation set was 0·79 (95% CI 0·77–0·81). The one-field AUC was 0·78 (0·76–0·80) in the internal validation set and 0·70 (0·67–0·74) in the external validation set (appendix p 15). In both validation sets, the AUC in eyes that had risk factor information available was similar to the AUC in all eyes (figure 1 A–C). Both the three-field and one-field versions of the deep-learning system were well calibrated for the internal validation set (appendix p 21). However, in the external validation set the one-field system overestimated the rate of developing diabetic retinopathy, which was resolved after scaling using 5% of the external validation set (appendix pp 3–4, 21). This scaling did not affect any other analysis.

Positive predictive values and negative predictive values were reported as a function of percentiles of risk threshold (figure 1D–F). Because the incidence of diabetic retinopathy was below 20%, the negative predictive values were generally higher than 80% and exceeded 95% for patients in the lowest-risk group. The positive predictive values increased sharply in patients at the highest risk, with the predicted risk of developing diabetic retinopathy ranging from 40% to 60%.

Univariable analyses compared the prognostication of the one field and three field deep-learning systems with assessment based on known risk factors; multivariable analysis evaluated whether the system could improve prognostication when added to risk factors for the internal validation set (table 2; figure 1A, B). The most predictive risk factor for the development of diabetic retinopathy in the univariable analyses was glycated haemoglobin (table 2). The combination of all risk factors gave an AUC of 0·72 (0·68–0·76). Addition of the three-field deep-learning system to the risk factors significantly increased the AUC to 0·81 (0·77–0·84; p<0·0001). The effect of adding the one-field deep-learning system was similar (table 2). The results of several weakly prognostic risk factors are reported in the appendix (p 10). The only overlapping risk factor between the external validation set and the development set that indicated the patient would develop retinopathy was glycated haemoglobin. Hypertension data were rarely available and weakly prognostic (appendix p 14). Therefore, we compared glycated haemoglobin alone to a model that included both glycated haemoglobin and the one-field deep-learning system. The



addition of the deep-learning system increased the AUC compared with risk factors alone in the external validation set (table 2 and figure 1C; p<0.0001). The difference between using the deep-learning system in isolation and adding risk factors to the system was much smaller, with absolute improvement ranging from 0·01 to 0·02 in both validation sets (table 2).

Kaplan-Meier curves illustrate the rate of development of diabetic retinopathy in the high-risk, medium-risk, and low-risk groups—defined using the upper and lower quartiles of deep-learning system prediction in the tuning dataset—predicted by the three-field deep-learning system for the internal validation set (figure 2A) and for the one-field deep-learning system assessment of the internal validation (figure 2B) and the external validation (figure 2C) sets. The rate of diabetic retinopathy development was significantly higher in the high-risk group than the low-risk group (p<0·0001). There were significant differences between the high-risk and the low-risk groups progressing to moderate or worse diabetic retinopathy (appendix p 22) and vision-threatening diabetic retinopathy (appendix p 22).

Explanation techniques were used to gain insight into the possible prognostic features of colour fundus photographs identified by the deep-learning system (figure 3; appendix p 17). Heatmaps of baseline fundus photographs were compared with follow-up fundus photographs. In some instances, the highlighted areas developed lesions later (figure 3A, B), whereas in other cases the most prominent highlighted areas did not develop lesions (figure 3C). For cases that already had subtle microaneurysms, the heatmaps highlighted the existing lesions (figure 3D).

We also assessed input ablation to find the most important image regions for accurate prognosis prediction. To do this one or two fields of the three-field deep-learning system were removed. By removing two of the three fields, the highest AUC was obtained using the primary field; however, when one field was removed, the highest AUC was obtained when the fields covered the widest field of view, with nasal and temporal inputs (figure 4A, B; appendix p 4). In the assessment of the one-field deep-learning system, the primary field was split into five regions for inclusion or exclusion (figure 4C, D). The central region alone resulted in the highest AUC, with the inferior region alone giving the lowest AUC.

Figure 2: **Kaplan-Meier analysis of the incidence of mild or worse diabetic retinopathy**
Incidence of diabetic retinopathy for the three-field DLS for the internal validation set (A), one-field DLS for the internal validation set (B), and one-field DLS for the external validation set (C). All Kaplan-Meier curves had log-rank test p<0·0001. Shaded areas are 95% CIs. Additional plots for incidence of moderate or worse diabetic retinopathy and vision-threatening diabetic retinopathy are presented in appendix (p 22). DLS=deep-learning system. HR=hazard ratio.

Discussion

We created two deep-learning systems to predict the development of diabetic retinopathy within 2 years, and validated them on two datasets: an internal validation set containing images from predominantly Hispanic patients from the USA, and an external validation set from Thailand. On both datasets, the deep-learning system had good performance both in isolation, and when adjusted for available risk factors. When combined with available risk factors, the prognostication improved compared with using the risk factors alone. Kaplan-Meier analyses showed that the deep-learning system's prognostication generalised to predicting incident diabetic retinopathy beyond 2 years and predicting moderate diabetic retinopathy and vision-threatening diabetic retinopathy. The differences in calibration between the two validation sets are discussed in the appendix (pp 3–4).

Several algorithms for stratifying diabetic retinopathy risk have been described, such as using individual risk factors to reduce screening frequency,[19] using microaneurysm turnover rate and central macular thickness to predict progression to diabetic macular oedema,[20] and using retinal arteriolar dilation to predict incident diabetic retinopathy.[21] Multifocal electroretinogram was also shown to predict new retinopathy development at specific retinal locations.[22] Additionally, deep learning was applied to colour fundus photographs to predict progression by two or more steps on the Early Treatment Diabetic Retinopathy Study scale.[22] The limitations of the study by Arcadu and colleagues[22] were the absence of adjusted analysis for risk factors, absence of an external validation set, small study size (530 patients) and the consequent use of cross-validation, and restrictive inclusion criteria (patients from two clinical trials).

Our study improves upon previous work in several ways. First, we consider the challenging task of

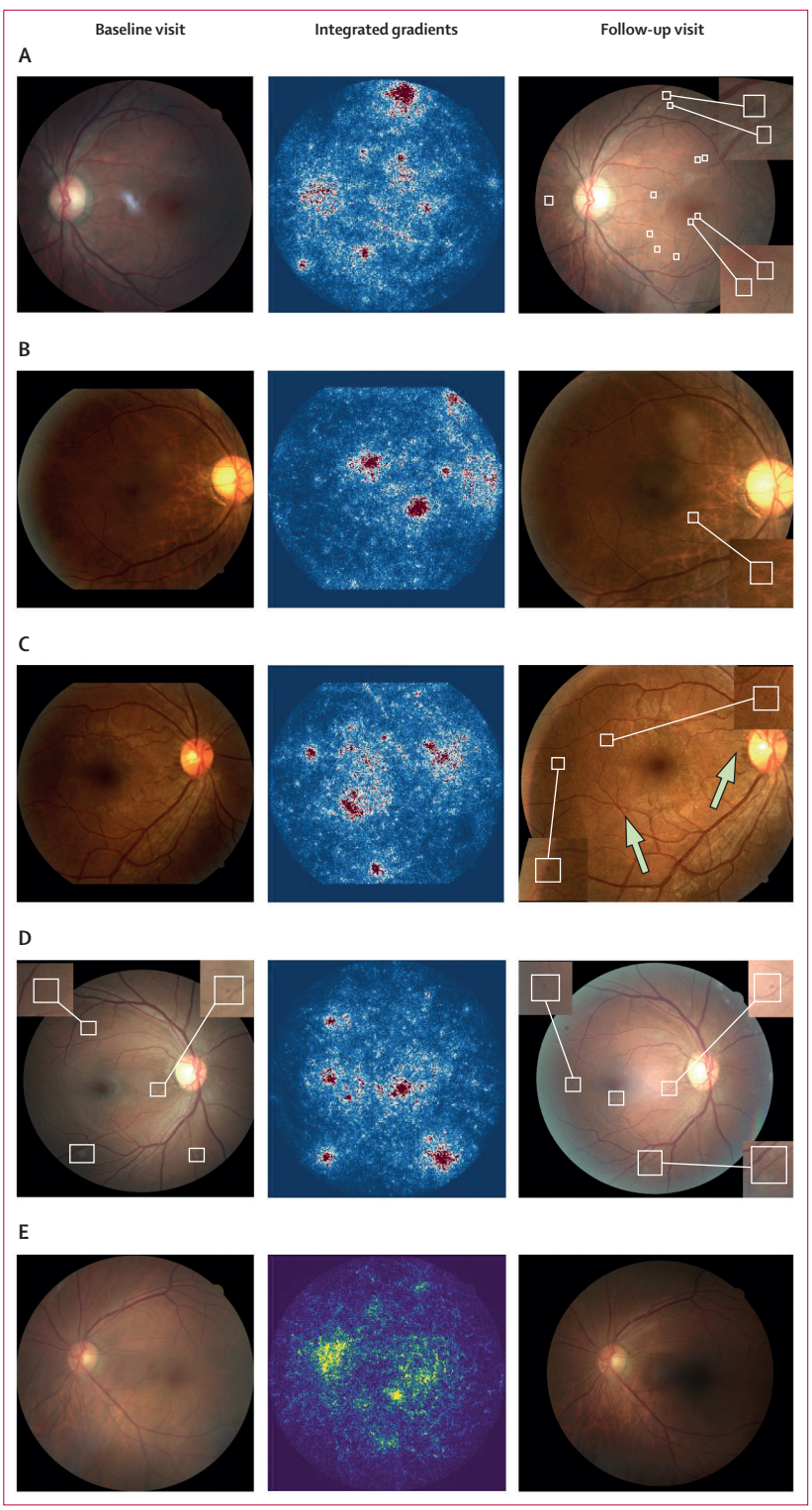


Figure 3: **Comparisons of colour fundus photographs at baseline and at follow-up using saliency heatmaps**

In each integrated gradient[18] saliency heatmap, the colour represents the strength of the contribution towards the prediction; red indicates a contribution towards developing diabetic retinopathy, yellow indicates contribution towards not developing diabetic retinopathy, whereas blue indicates little contribution. Microaneurysms are highlighted in both baseline and follow-up images. (A) This 43-year-old patient did not have diabetic retinopathy at baseline visit, but developed diabetic retinopathy on a future follow-up visit when aged 44 years. The prediction of developing diabetic retinopathy was high (0·75) and the highlighted regions in the saliency map eventually developed microaneurysms. Note that not all future microaneurysms were highlighted. (B) This 48-year-old patient did not have diabetic retinopathy at baseline, but developed diabetic retinopathy on a future follow-up visit when aged 50 years. The prediction was high (0·71) and the highlighted regions in the saliency map eventually developed microaneurysms. (C) This 34-year-old patient did not have diabetic retinopathy at the baseline, but developed diabetic retinopathy at a follow-up visit when aged 37 years. The prediction was high (0·68), but the most notable highlighted regions (indicated by green arrows) did not show microaneurysms at the follow-up visit. (D) Patient had a very high DLS prediction (0·97), and upon closer inspection the baseline image showed subtle signs of mild diabetic retinopathy, as evidenced by microaneurysms. The saliency map correctly highlighted the microaneurysms at baseline. (E) This 64-year-old patient did not have diabetic retinopathy at baseline and did not develop diabetic retinopathy at the latest available follow-up visit when aged 70 years. The DLS correctly assigned a small probability of progression (0·06; appendix p 4). DLS=deep-learning system.

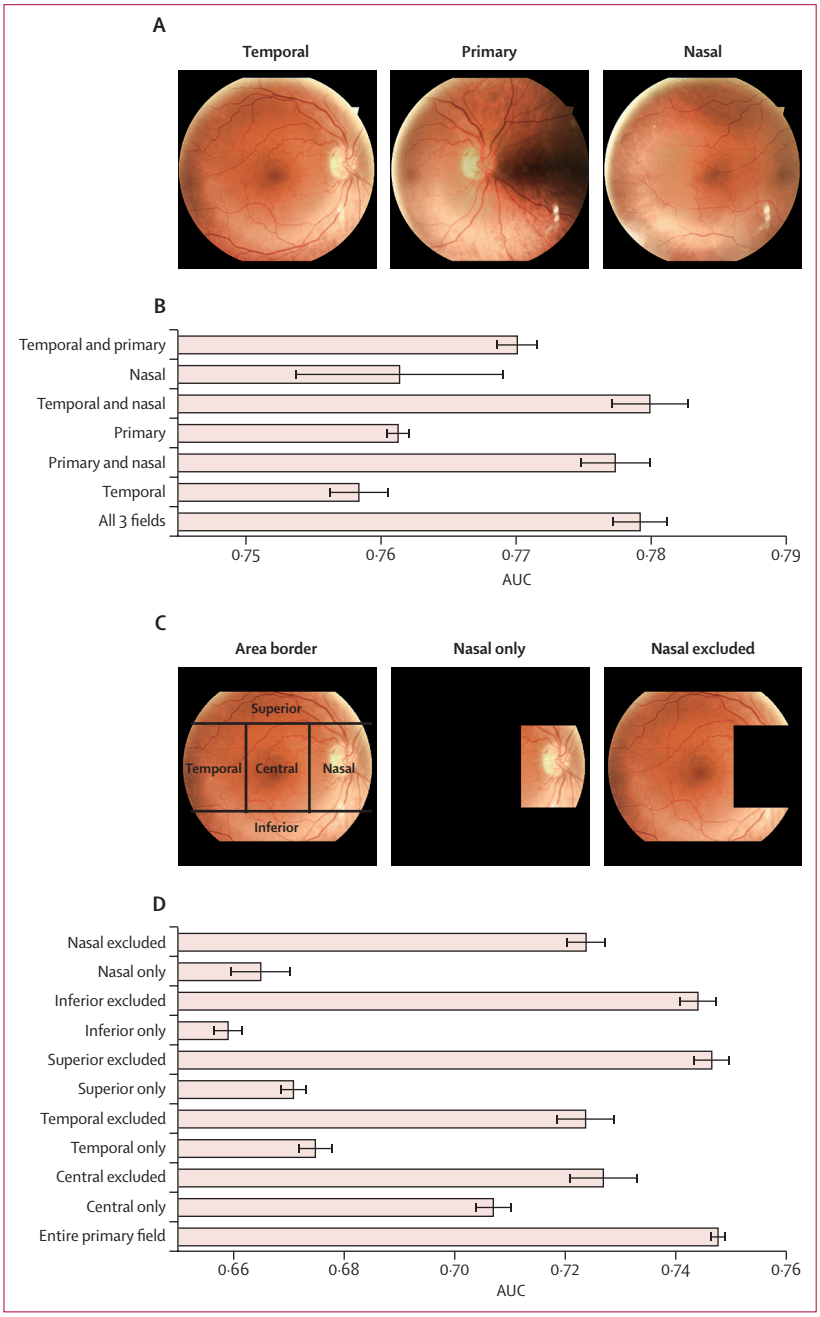


Figure 4: **Assessment of the different fields and parts of images of the eye**
(A) Sample images of the temporal, primary, and nasal fields. (B) DLS performance using one or two images from three-field images compared with the performance using all three images. (C) The left column contains sample images of the primary field with several different regions. (D) DLS performance after systematically including only each region or excluding that region, compared with the performance using all the regions. Error bars similarly indicate one standard deviation across three training runs. AUC=area under the curve. DLS=deep-learning system.

stratifying patients by their risk of developing diabetic retinopathy with colour fundus photographs and risk factors, which are available in most screening settings. This approach directly tackles the problem of optimising screening intervals by stratifying the largest group of patients: those without any diabetic retinopathy at baseline. Second, our system retained low prognostic value after adjusting for available risk factors. Third, we validated our system on two separate validation datasets in two countries from two continents. Our system retained substantial predictive value across both validation sets despite differences in patient populations, fundus cameras (appendix p 1), glycated haemoglobin concentrations (appendix p 14), average incidences (table 1), and grading protocols (appendix pp 6–7). In addition, our development sets spanning multiple years of follow-up are orders of magnitude larger than previous work by Arcadu and colleagues.[23] Finally, the risk groups constructed based on our system retained statistically significant separations in terms of the incidence of diabetic retinopathy even when evaluated against endpoints corresponding to different severities of diabetic retinopathy (appendix pp 22).

In addition to formulation of the problem and validation of results, we did several analyses to better understand prognostic features of the colour fundus photographs. First, our deep-learning system predicts a high likelihood of developing diabetic retinopathy when subtle microaneurysms are present at baseline. Although these eyes were meant to be excluded from the study, subtle microaneurysms can be missed in real-world settings; as such, it is reassuring that the deep-learning system identifies such patients, which could prompt graders to examine the case more carefully. Second, the heatmap analysis suggests that the deep-learning system sometimes finds the regions that eventually develop diabetic retinopathy to be of most importance. This finding suggests the presence of subtle signs that are not visually apparent, a phenomenon that merits study. Third, the primary field was found to be the most important for the prediction of incident diabetic retinopathy, but the best combination of fields was the nasal and temporal (excluding the primary field; figure 4). These observations can be reconciled by considering that the combination of the nasal and temporal fields encompasses the primary field, and that these two fields when combined provide a more expansive view of the retina. Finally, within the primary field, the superior and inferior regions had the smallest effect on deep-learning system prognostic ability when excluded, suggesting that they were the least important. Conversely, the macular region had the highest effect when removed and was also the most prognostic in isolation. Thus, the deep-learning system attributed high importance to both the macula and the periphery of the retina in prognostication.

One application of this deep-learning system might be the optimisation of screening intervals. Effective treatments, such as laser photocoagulation and intravitreal injections of antivascular endothelial growth factor,[24–26] are more effective the earlier the retinopathy is detected. Although diabetic retinopathy screening traditionally relied on direct or indirect ophthalmoscopy or slit lamp biomicroscopy, the ease-of-use, cost-effectiveness, and

accuracy[27] of fundus photography has merited its recommendation in multiple diabetic retinopathy screening guidelines.[3,4,28] The range of screening intervals in these guidelines (eg, 12–24 months for patients with no apparent diabetic retinopathy) accounts for clinical risk factors, resource availability (eg, imaging equipment, supplies, and personnel), and other socioeconomic factors. Our deep-learning system can provide a more accurate personalised risk assessment to optimise screening intervals; patients at high risk can be followed up frequently to ensure early detection, whereas patients at low risk could be followed up less frequently to reduce the screening burden shared by patients, clinicians, and the health-care system. In our study the patients at the highest risk developed diabetic retinopathy at a rate exceeding 80%, whereas patients at the lowest risk had a less than 5% chance of developing diabetic retinopathy. However, the specific cutoffs for defining high or low risk status and corresponding interventions will need to be studied in future work and probably be tailored to the local resource availability and practice patterns.

Other potential uses of the deep-learning system revolve around targeted interventions. For example, patients at high risk could be preferentially selected for more intensive lifestyle modifications or counselling.[29,30] These patients might be better suited for stricter pharmacological control of blood sugar or as candidates in relevant clinical trials. Patients sometimes miss their screening visits, and, in such cases, our deep-learning system could be used to identify and alert (eg, via telephone or texts) patients at high risk of developing diabetic retinopathy to improve screening compliance and consequently visual outcomes. Finally, the deep-learning system's ability to predict diabetic retinopathy development beyond 2 years suggests its usefulness in long-term population-level forecasting for public health planning.

Our study has several limitations. First, although glycated haemoglobin data were available for both datasets, several known risk factors or comorbidities, including blood pressure measurements, were not. Some risk factors, such as glycated haemoglobin, were also potentially from an earlier blood test instead of a test at the time of diabetic retinopathy screening. Similarly, device information was not available on a per-image basis; thus, we could not evaluate deep-learning system performance stratified by imaging device. Second grading variability exists, especially for subtle findings, such as microaneurysms, resulting in some patients already having mild diabetic retinopathy at baseline. Although the deep-learning system prediction that these patients would develop diabetic retinopathy is semantically correct, users should be aware that a high predicted risk might indicate existing diabetic retinopathy. Potential solutions to this include the concurrent use of a diabetic retinopathy grading algorithm with this risk stratification tool, which could be investigated in future work. Similarly, whether the deep-learning system can help to prognosticate patients with existing diabetic retinopathy (whether mild or on more granular grading scales) could be studied. Third, because our diabetic retinopathy grades were based on one-field or three-field colour fundus photographs, lesions outside of these fields might remain undetected and the absence of optical coherence tomography might reduce diabetic macular oedema detection accuracy. Since the data came from screening programmes, patients in both datasets were generally referred for ophthalmology follow-up when moderate or worse diabetic retinopathy was discovered, rendering the evaluation of progression to vision-threatening diabetic retinopathy non-ideal. Finally, our evaluation was on a single randomly selected eye per patient in retrospective settings. A patient-level analysis, ideally under prospective settings, will help to evaluate the clinical relevance.

Our results suggest that a deep-learning system could be developed to improve risk stratification for developing diabetic retinopathy. Future studies could increase the effectiveness of interventions tailored to progression risk (eg, screening intervals) towards improving health outcomes and reducing cost.

Contributors

AB and PB verified the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. AB, SB, SVi, AVV, YL, and PB conceived and designed the study. AB, SB, BB, SVe, AM, NH, YL, and PB analysed the data. AB, GOM, JC, and PR were involved in data collection. GSC, LP, DRW, and AVV provided strategic guidance and oversight. AB, SB, NH, and YL drafted the manuscript with input from all authors. The final version of the Article has been approved by all the authors.

Declaration of interests

LP, DRW, AV, and YL have a patent on processing fundus images using machine learning models (US20190180441A1), and are employees of Google and own stock in Alphabet. GSC has patents on machine learning for medical applications (US 20170032241A1; US10402721B2), is an employee of Google and owns stock in Alphabet. SVi has provisional patent filings in medical imaging. AB, BB, SVi, SVe, AM, GOM, GSC, NH, and PB are Google employees and own stock in Alphabet. SB was a paid consultant for Google. JC is the chief executive officer of EyePACS. PR received grants and personal fees from Novartis and Roche and personal fees from Bayer.

Data sharing

De-identified data used in this study are not publicly available at present. Parties interested in data access should contact JC (jcuadros@eyepacs.com) for queries related to EyePACS, and PR (paisan.trs@gmail.com) for queries related to the Thailand dataset. Applications will need to undergo ethical and legal approvals by the respective institutions. Those interested in research collaborations should contact NH (nhammel@google.com). The deep learning algorithm can be trained via the publicly available Inception-v3 architecture starting from the pretrained network.

Acknowledgments

This study was funded by and undertaken by employees of Google. The authors would like to thank Jacqueline Shreibati, Ellery Wulczyn, and Michael Howell for reviewing and suggesting improvements to the manuscript; and Roy Lee, Noemi Figueroa, and the labelling software team in Google Health for assistance in data labelling.

References

1 Cheung N, Mitchell P, Wong TY. Diabetic retinopathy. *Lancet* 2010; **376:** 124–36.

For more on the **Inception-v3 architecture** see https://github.com/tensorflow/models/blob/master/research/slim/nets/inception_v3.py

For more on the **pretrained network** see https://github.com/tensorflow/models

2 Solomon SD, Chew E, Duh EJ, et al. Diabetic retinopathy: a position statement by the American Diabetes Association. *Diabetes Care* 2017; **40:** 412–18.

3 International Council of Ophthalmology. ICO guidelines for diabetic eye care. 2017. http://www.icoph.org/downloads/ICOGuidelinesforDiabeticEyeCare.pdf (accessed June 16, 2020).

4 American Academy Of Ophthalmology. Diabetic retinopathy preferred practice patterns. 2019 https://www.aao.org/preferred-practice-pattern/diabetic-retinopathy-ppp (accessed June 16, 2020).

5 Ogurtsova K, da Rocha Fernandes JD, Huang Y, et al. IDF diabetes atlas: global estimates for the prevalence of diabetes for 2015 and 2040. *Diabetes Res Clin Pract* 2017; **128:** 40–50.

6 Scanlon PH. Screening intervals for diabetic retinopathy and implications for care. *Curr Diab Rep* 2017; **17**: 96.

7 Jenkins AJ, Joglekar MV, Hardikar AA, Keech AC, O'Neal DN, Januszewski AS. Biomarkers in diabetic retinopathy. *Rev Diabet Stud* 2015; **12:** 159–95.

8 Stratton IM, Kohner EM, Aldington SJ, et al. UKPDS 50: risk factors for incidence and progression of retinopathy in type II diabetes over 6 years from diagnosis. *Diabetologia* 2001; **44:** 156–63.

9 Grading diabetic retinopathy from stereoscopic color fundus photographs—an extension of the modified Airlie House classification. ETDRS report number 10. Early Treatment Diabetic Retinopathy Study Research Group. *Ophthalmology* 1991; **98:** 786–806.

10 Wilkinson CP, Ferris FL 3rd, Klein RE, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology* 2003; **110:** 1677–82.

11 The Royal College of Ophthalmologists. The Royal College of Ophthalmologists diabetic retinopathy guidelines, 2013. https://www.rcophth.ac.uk/wp-content/uploads/2014/12/2013-SCI-301-FINAL-DR-GUIDELINES-DEC-2012-updated-July-2013.pdf (accessed June 16, 2020).

12 Jones CD, Greenwood RH, Misra A, Bachmann MO. Incidence and progression of diabetic retinopathy during 17 years of a population-based screening program in England. *Diabetes Care* 2012; **35:** 592–96.

13 Modjtahedi BS, Theophanous C, Chiu S, Luong TQ, Nguyen N, Fong DS. Two-year incidence of retinal intervention in patients with minimal or no diabetic retinopathy on telemedicine screening. *JAMA Ophthalmol* 2019; **137:** 445.

14 Ruamviboonsuk P, Krause J, Chotcomwongse P, et al. Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. *NJP Digital Med* 2019; **2:** 25.

15 US Department of Health & Human Services. The HIPAA Privacy Rule. 2008. https://www.hhs.gov/hipaa/for-professionals/privacy/index.html (accessed Sept 11, 2020).

16 Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. The IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA; June 27–30, 2016 (abstr).

17 Google. Google-research/tf-slim. 2019. https://github.com/google-research/tf-slim (accessed June 17, 2020).

18 Xu S, Venugopalan S, Sundararajan M. Attribution in scale and space. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit* 2020; 9680–89

19 Estil S, Steinarsson AÞ, Einarsson S, Aspelund T, Stefánsson E. Diabetic eye screening with variable screening intervals based on individual risk factors is safe and effective in ophthalmic practice. *Acta Ophthalmol* 2020; **98:** 343–46.

20 Cunha-Vaz J, Ribeiro L, Costa M, Simó R. Diabetic retinopathy phenotypes of progression to macular edema: pooled analysis from independent longitudinal studies of up to 2 years' duration. *Invest Ophthalmol Vis Sci* 2017; **58:** BIO206–10.

21 Rogers SL, Tikellis G, Cheung N, et al. Retinal arteriolar caliber predicts incident retinopathy: the Australian diabetes, obesity and lifestyle (AusDiab) study. *Diabetes Care* 2008; **31:** 761–63.

22 Bearse MA Jr, Adams AJ, Han Y, *et al*. A multifocal electroretinogram model predicting the development of diabetic retinopathy. *Prog Retin Eye Res* 2006; **25:** 425–48.

23 Arcadu F, Benmansour F, Maunz A, Willis J, Haskova Z, Prunotto M. Deep learning algorithm predicts diabetic retinopathy progression in individual patients. *NPJ Digit Med* 2019; **2:** 92.

24 Rohan TE, Frost CD, Wald NJ. Prevention of blindness by screening for diabetic retinopathy: a quantitative assessment. *BMJ* 1989; **299:** 1198–201.

25 Ferris FL 3rd. How effective are treatments for diabetic retinopathy? *JAMA* 1993; **269:** 1290–91.

26 Zhao Y, Singh RP. The role of anti-vascular endothelial growth factor (anti-VEGF) in the management of proliferative diabetic retinopathy. *Drugs Context* 2018; **7:** 212532.

27 Lin DY, Blumenkranz MS, Brothers RJ, Grosvenor DM. The sensitivity and specificity of single-field nonmydriatic monochromatic digital fundus photography with remote image interpretation for diabetic retinopathy screening: a comparison with ophthalmoscopy and standardized mydriatic color photography. *Am J Ophthalmol* 2002; **134:** 204–13.

28 American Diabetes Association. Microvascular complications and foot care: standards of medical care in diabetes–2020. *Diabetes Care* 2020; **43** (suppl 1)**:** S135–51.

29 Dixon RF, Zisser H, Layne JE, et al. A virtual type 2 diabetes clinic using continuous glucose monitoring and endocrinology visits. *J Diabetes Sci Technol* 2019; **14:** 908–11.

30 Downing J, Bollyky J, Schneider J. Use of a connected glucose meter and certified diabetes educator coaching to decrease the likelihood of abnormal blood glucose excursions: the livongo for diabetes program. *J Med Internet Res* 2017; **19:** e234.