







Detection of signs of disease in external photographs of the eyes via deep learning

Boris Babenko^{1,7}, Akinori Mitani^{1,2,7}, Ilana Traynis³, Naho Kitade¹, Preeti Singh¹, April Y. Maa^{4,5}, Jorge Cuadros⁶, Greg S. Corrado¹, Lily Peng¹, Dale R. Webster¹, Avinash Varadarajan¹, Naama Hammel¹   and Yun Liu¹  

Retinal fundus photographs can be used to detect a range of retinal conditions. Here we show that deep-learning models trained instead on external photographs of the eyes can be used to detect diabetic retinopathy (DR), diabetic macular oedema and poor blood glucose control. We developed the models using eye photographs from 145,832 patients with diabetes from 301 DR screening sites and evaluated the models on four tasks and four validation datasets with a total of 48,644 patients from 198 additional screening sites. For all four tasks, the predictive performance of the deep-learning models was significantly higher than the performance of logistic regression models using self-reported demographic and medical history data, and the predictions generalized to patients with dilated pupils, to patients from a different DR screening programme and to a general eye care programme that included diabetics and non-diabetics. We also explored the use of the deep-learning models for the detection of elevated lipid levels. The utility of external eye photographs for the diagnosis and management of diseases should be further validated with images from different cameras and patient populations.

Disease diagnosis and management often require specialized equipment and trained medical professionals to interpret findings. For example, diabetic retinopathy (DR) screening programmes use either a fundoscope or a fundus camera to examine the posterior part of the eye (that is, the retinal fundus; Fig. 1a). Such examinations can reveal diabetes-related blood vessel compromise, such as microaneurysms. More recently, machine-learning-based assessments of retinal fundus photographs have been shown to detect signs of cardiovascular risk^{1,2}, anaemia³, chronic kidney disease⁴ and other systemic parameters⁵. Despite the expanding diagnostic information that can be obtained from fundus photographs, the burden of costly specialized cameras, skilled imaging technicians and, oftentimes, mydriatic eye drops to dilate the patients' pupils limits the use of these diagnostic techniques to eye clinics or primary care facilities with specialized equipment.

By contrast, the anterior part of the eye can be readily imaged to produce external eye photographs (Fig. 1a) without specialized equipment or mydriatic eye drops. Indeed, external eye photographs have an associated Current Procedural Terminology code (92285) and are sometimes used to document progression of anterior eye conditions, such as eyelid abnormalities, corneal ulcers and cataracts. Interestingly, the anterior part of the eye is also known to manifest signs of disease beyond those affecting the front of the eye. For example, hypertension can cause recurrent subconjunctival haemorrhages (from broken blood vessels in the white part of the eye)⁶, and diabetes can affect pupil size^{7,8}, conjunctival vessel calibre⁹ and tortuosity¹⁰.

Here we hypothesized that deep learning can more precisely detect disease presence from external eye photographs and potentially extract information that can help in identifying high-risk individuals for further testing or follow-up. To test this hypothesis,

we leveraged de-identified data from tele-ophthalmology screening programmes where important parameters of systemic health and diabetic eye care were available, including glycated haemoglobin (HbA1c, a marker of glucose control), total cholesterol, triglycerides and diabetic retinal disease diagnoses (DR, diabetic macular oedema (edema) (DME) and vision-threatening DR (VTDR, a combination of DR and DME)). Our results indicate that external eye photographs can indeed reveal signs of systemic and retinal disease and merit further study.

Results

We developed a deep-learning system (DLS) using external eye images taken with fundus cameras (Fig. 1a) from 145,832 patients with diabetes screened for DR at 301 sites in California, USA (Fig. 1b). To evaluate the DLS, we used four validation datasets (Table 1). The first validation set ('A') included 27,415 patients who underwent DR screening at 186 sites across 18 US states without pupillary dilation (that is, without use of mydriatic eye drops); the second validation set ('B') included 5,058 patients at 59 sites across 14 US states with pupillary dilation. The third validation set ('C') included 10,402 patients who underwent mydriatic DR screening at the Atlanta Veterans Affairs (VA) hospital in Georgia, USA as part of the diabetic teleretinal screening programme. The final validation set ('D') included 6,266 patients from the Technology-based Eye Care Services (TECS) programme in Georgia, USA. In this programme, screening was mydriatic and located in primary care clinics surrounding the Atlanta VA hospital, and the screened population included patients both with and without diabetes. The pre-specified primary analyses included four predictions: poor blood glucose control (defined as HbA1c $\geq 9\%$ by the Healthcare Effectiveness Data and Information Set (HEDIS¹¹)), moderate-or-worse ('moderate+')

¹Google Health, Palo Alto, CA, USA. ²Artera, Mountain View, CA, USA. ³Google Health via Advanced Clinical, Deerfield, IL, USA. ⁴Department of Ophthalmology, Emory University School of Medicine, Atlanta, GA, USA. ⁵Regional Telehealth Services, Technology-based Eye Care Services (TECS) Division, Veterans Integrated Service Network (VISN) 7, Decatur, GA, USA. ⁶EyePACS Inc, Santa Cruz, CA, USA. ⁷These authors contributed equally: Boris Babenko, Akinori Mitani. [✉]e-mail: nhammel@google.com; liuyun@google.com



Fig. 1 | Extracting insights from external photographs of the front of the eye. **a**, Diabetes-related complications can be diagnosed by using specialized cameras to take fundus photographs, which visualize the posterior segment of the eye. By contrast, anterior imaging using a standard consumer-grade camera can reveal conditions affecting the eyelids, conjunctiva, cornea and lens. In this work, we show that external photographs of the eye can offer insights into diabetic retinal disease and detect poor blood sugar control. These images are shown for illustrative purposes only and do not necessarily belong to the same subject. **b**, Our external eye DLS was developed on data from California (CA, yellow) and evaluated on data from 18 other US states (light blue). **c**, AUCs for the four validation datasets. The prediction tasks shown are poor sugar control (HbA1c $\geq 9\%$), elevated lipids (total cholesterol (Tch) ≥ 240 mg dl⁻¹, triglycerides (Tg) ≥ 200 mg dl⁻¹), moderate+ DR, DME, VTDR and a positive control: cataract. Some targets were not available in validation sets C and D. Error bars are 95% CIs computed using the DeLong method. A graphic visualization of receiver operating characteristic (ROC) curves is shown in Extended Data Fig. 3, and Supplementary Table 1 contains numerical values of AUCs along *P* values for improvement. **d**, Similar to **c** but for PPV. The thresholds used for the PPV are based on the 5% of patients with the highest predicted likelihood. Error bars are 95% bootstrap CIs. A graphic visualization of the PPV over multiple thresholds is in Extended Data Fig. 1. ^aBaseline characteristics models for validation sets A and B include self-reported age, sex, race/ethnicity and years with diabetes and were trained on the training dataset. ^bThe baseline characteristics models for validation sets C and D include self-reported age and sex and were trained directly on the respective sets due to large differences in patient population compared with the development set. Thus, these overestimate baseline performance. ^cPre-specified primary prediction tasks.

Table 1 | Characteristics of the datasets

Datasets	Development set		Validation sets				
	Training set	Tuning set	A	B	C	D	
Source	EyePACS (CA)		EyePACS (non-CA)		VA	VA	
Number of US states	1 (CA)	1 (CA)	18 (non-CA)	14 (non-CA)	1 (GA)	1 (GA)	
Number of sites	277 ^a	54 ^a	186 ^a	59 ^a	9	6	
Number of patients	126,066	19,766	27,415	5,058	10,402	6,266	
Number of visits ^b	159,269	23,095	27,415	5,058	10,402	6,266	
Number of images ^b	290,642	41,928	53,861	9,853	19,763	12,751	
Dilation status (%)	No	137,710 (47%)	21,063 (50%)	27,415 (100%)	0 (0%)	150 (1%)	NA
	Yes	100,929 (35%)	9,678 (23%)	0 (0%)	5,058 (100%)	9,788 (94%)	NA
	Unknown	52,003 (18%)	11,187 (27%)	0 (0%)	0 (0%)	464 (4%)	6,266 (100%)
Age (years, mean \pm s.d.)	54 \pm 11	54 \pm 11	54 \pm 12	53 \pm 11	63 \pm 10	62 \pm 13	
Self-reported sex (%)	Female	72,880 (58%)	11,655 (59%)	14,525 (53%)	3,004 (59%)	478 (5%)	944 (15%)
	Male	53,186 (42%)	8,111 (41%)	12,890 (47%)	2,054 (41%)	9,924 (95%)	5,322 (85%)
Race/ethnicity (%)	Hispanic	97,300 (77%)	15,577 (79%)	11,393 (42%)	4,152 (82%)	-	0 (0%)
	White	11,459 (9%)	1,609 (8%)	7,621 (28%)	374 (7%)	45% ^c	2,577 (43%)
	Black	4,991 (4%)	1,074 (5%)	4,140 (15%)	398 (8%)	49% ^c	3,317 (56%)
	Asian/Pacific islander	7,954 (6%)	1,091 (6%)	2,949 (11%)	62 (1%)	<1%	38 (1%)
	Native American	1,669 (1%)	83 (0%)	411 (1%)	37 (1%)	<1%	36 (1%)
	Other	2,693 (2%)	332 (2%)	901 (3%)	35 (1%)	-	0 (0%)
Years with diabetes (years, median (interquartile range))	5 (2-13)	8 (2-13)	5 (2-8)	8 (2-13)	NA	NA	

^aThirty sites overlapped between the training set and the tuning set because patients who visited both training set sites and tuning set sites were randomly assigned to either set. Fifty-six sites overlapped between validation sets A and B because they included both patients with dilated and non-dilated pupils. There was no patient-level overlap between the development set and the validation sets. ^bIn the development set, all visits, eyes and images were used. In each validation set, one random visit was selected per patient. Although all evaluations were performed on the patient level, for eye disease evaluation, one random eye per patient was selected and, for HbA1c, evaluation images of both eyes were used (and the DLS predictions averaged). Note that not all labels were available in some visits, and the number of visits/images used for evaluating each task can be smaller than the numbers here. See Tables 2–5 for the number of visits used for the evaluation of each task. ^cPatient-level information not available; these reflect a cohort-level estimate and are consistent with the predominantly Black/White population in previous studies⁵⁸. CA, California; GA, Georgia; NA, not applicable; -, data not available.

DR, DME and VTDR. Cataract detection was a fifth prediction task used as a positive control. In addition, we present exploratory analyses for predictions of elevated total cholesterol and triglycerides relative to several clinical cut-offs.

We first evaluated the DLS for its ability to detect HbA1c $\geq 9\%$. Across the four validation sets A to D, the DLS achieved an area under the receiver operating characteristic curve (AUC) of 70.0%, 73.4%, 69.3% and 67.6%, respectively (Fig. 1c and Supplementary Table 1; full ROC curves shown in Extended Data Fig. 3). For comparison, logistic regression models using baseline characteristics (age, sex, race/ethnicity, years with diabetes on validation sets A and B, age and sex on validation sets C and D) achieved AUCs of 64.8%, 66.5%, 65.2% and 59.8%, respectively ($P < 0.001$ for comparing the DLS with baseline characteristics alone in all three validation sets). The AUC trends for predicting other thresholds (HbA1c $\geq 7\%$ and $\geq 8\%$) were similar and are presented in Supplementary Table 1. The positive predictive values (PPVs) for HbA1c $\geq 9\%$ exceeded 69% for the 5% of patients with the highest risk predicted by the DLS, and the negative predictive values (NPVs) exceeded 88% for the lowest-risk 5% in validation set A (Fig. 1d and Extended Data Figs. 1a and 2a).

Next, we evaluated the DLS for detecting diabetic retinal diseases (as graded by certified teleretinal experts using fundus photographs; see 'Labels' section in Methods). The AUCs for the three conditions (moderate+ DR, DME and VTDR) were 75.0%, 77.9% and 79.2%, respectively, in validation set A. On dilated pupils, the AUCs were consistently 5–10% higher, at 84.0%, 84.7% and 86.7% in validation set B, and more similar to validation set A, at 76.8%, 79.5% and 81.1%, in validation set C (which was from a different patient population). Using only baseline characteristics resulted in AUCs

that were 4–10% lower in validation sets A and B and 13–24% lower in validation set C ($P < 0.001$ for comparing the DLS with baseline characteristics alone for all comparisons). The trends for other DR thresholds ('mild+' and 'severe+') were similar and are presented in Supplementary Table 1. The PPV and NPV for the various diabetic retinal diseases similarly indicated the ability to identify patients at either very high or very low likelihood of these conditions (Fig. 1d and Extended Data Figs. 1d–f and 2d–f).

Finally, we evaluated the DLS for detecting elevations in available elements of the lipid panel: total cholesterol and triglycerides. Across the four validation datasets, the AUCs for total cholesterol ≥ 240 mg dl⁻¹ were 59.1% (A), 57.9% (B), 62.3% (C) and 58.1% (D), and for triglycerides the AUCs for ≥ 200 mg dl⁻¹ were 63.6% (A), 62.7% (B), 66.3% (C) and 67.1% (D) (Supplementary Table 1), with similar trends for other cut-offs (total cholesterol ≥ 200 mg dl⁻¹, triglycerides ≥ 150 mg dl⁻¹ and ≥ 500 mg dl⁻¹). For this set of predictions, the DLS did not consistently outperform the baseline characteristics (see adjusted analyses in Tables 2–5 for complementary analyses).

Adjusted analyses. To verify that the DLS predictions were not solely mediated by baseline characteristics as confounders, we examined the odds ratios of the DLS predictions adjusted for baseline characteristics (Tables 2–5). For all primary prediction tasks in all four validation datasets, DLS predictions (standardized to zero mean and unit variance for interpretability of coefficients) had statistically significant adjusted odds ratios ($P < 0.001$). The adjusted odds ratios ranged from 1.5 to 2.0 among tasks for validation set A, from 2.0 to 3.1 for validation set B, from 1.6 to 2.0 for validation set C and it was 2.0 for validation set D. For elevated lipids, despite the fact that the

Table 2 | Adjusted analysis for validation set A

Prediction target	Number of visits		Odds ratio (95% CI), P value						
	Total	Positive (%)	Age (per decade)	Male	White ^a	Black ^a	Asian/Pacific islander ^a	Years with diabetes (per five years)	DLS (external eye image)
HbA1c ≥ 7%	21,183	13,544 (63.9%)	0.891	1.106	0.962	0.976	1.077	1.451	1.720
			(0.866–0.918)	(1.042–1.175)	(0.892–1.038)	(0.892–1.067)	(0.976–1.188)	(1.408–1.496)	(1.660–1.782)
			P < 0.001	P < 0.001	P = 0.314	P = 0.593	P = 0.140	P < 0.001	P < 0.001
HbA1c ≥ 8%	21,183	9,347 (44.1%)	0.916	1.073	0.907	0.925	1.029	1.326	1.955
			(0.889–0.944)	(1.011–1.138)	(0.843–0.977)	(0.847–1.010)	(0.932–1.135)	(1.290–1.363)	(1.883–2.030)
			P < 0.001	P = 0.019	P = 0.010	P = 0.083	P = 0.572	P < 0.001	P < 0.001
HbA1c ≥ 9% ^b	21,183	6,537 (30.9%)	0.934	1.072	0.911	0.965	0.944	1.210	1.971
			(0.904–0.965)	(1.007–1.142)	(0.842–0.986)	(0.878–1.060)	(0.847–1.051)	(1.176–1.245)	(1.895–2.051)
			P < 0.001	P = 0.030	P = 0.021	P = 0.459	P = 0.290	P < 0.001	P < 0.001
Tch ≥ 200 mg dl ⁻¹	9,318	2,913 (31.3%)	0.966	0.921	0.963	0.878	0.833	0.915	1.251
			(0.925–1.009)	(0.841–1.010)	(0.861–1.078)	(0.781–0.987)	(0.696–0.996)	(0.879–0.954)	(1.188–1.318)
			P = 0.122	P = 0.079	P = 0.517	P = 0.030	P = 0.045	P < 0.001	P < 0.001
Tch ≥ 240 mg dl ⁻¹	9,318	984 (10.6%)	0.981	0.946	1.065	0.858	0.808	0.916	1.328
			(0.921–1.046)	(0.827–1.083)	(0.901–1.259)	(0.717–1.028)	(0.606–1.079)	(0.861–0.975)	(1.241–1.420)
			P = 0.564	P = 0.423	P = 0.459	P = 0.096	P = 0.149	P = 0.006	P < 0.001
Tg ≥ 150 mg dl ⁻¹	9,104	4,674 (51.3%)	0.989	1.053	1.056	0.809	0.803	0.914	1.503
			(0.952–1.028)	(0.967–1.147)	(0.946–1.180)	(0.705–0.928)	(0.684–0.941)	(0.879–0.950)	(1.416–1.595)
			P = 0.589	P = 0.233	P = 0.330	P = 0.003	P = 0.007	P < 0.001	P < 0.001
Tg ≥ 200 mg dl ⁻¹	9,104	2,894 (31.8%)	0.939	1.144	1.028	0.717	0.773	0.935	1.433
			(0.900–0.980)	(1.044–1.254)	(0.915–1.155)	(0.620–0.830)	(0.646–0.926)	(0.897–0.975)	(1.348–1.522)
			P = 0.004	P = 0.004	P = 0.641	P < 0.001	P = 0.005	P = 0.002	P < 0.001
Tg ≥ 500 mg dl ⁻¹	9,104	323 (3.5%)	0.859	1.782	1.123	0.593	0.713	0.907	1.212
			(0.765–0.963)	(1.392–2.281)	(0.855–1.475)	(0.412–0.855)	(0.436–1.169)	(0.814–1.012)	(1.088–1.351)
			P = 0.009	P < 0.001	P = 0.405	P = 0.005	P = 0.180	P = 0.080	P < 0.001
Mild+DR	26,950	5,144 (19.1%)	0.901	1.160	0.781	1.047	1.301	1.641	1.796
			(0.876–0.928)	(1.086–1.240)	(0.718–0.850)	(0.952–1.152)	(1.165–1.454)	(1.595–1.688)	(1.741–1.853)
			P < 0.001	P < 0.001	P < 0.001	P = 0.342	P < 0.001	P < 0.001	P < 0.001
Moderate+DR ^b	26,950	3,247 (12.0%)	0.904	1.205	0.785	0.833	1.141	1.680	1.881
			(0.873–0.936)	(1.111–1.306)	(0.709–0.870)	(0.739–0.938)	(0.997–1.305)	(1.625–1.737)	(1.819–1.944)
			P < 0.001	P < 0.001	P < 0.001	P = 0.003	P = 0.056	P < 0.001	P < 0.001
Severe+DR	26,950	404 (1.5%)	0.870	1.516	0.803	1.084	2.216	1.840	1.444
			(0.798–0.949)	(1.234–1.863)	(0.612–1.052)	(0.795–1.477)	(1.651–2.975)	(1.695–1.999)	(1.384–1.507)
			P = 0.002	P < 0.001	P = 0.111	P = 0.611	P < 0.001	P < 0.001	P < 0.001
DME ^b	26,950	797 (3.0%)	0.899	1.241	0.633	0.989	1.206	1.664	1.475
			(0.843–0.958)	(1.072–1.438)	(0.519–0.772)	(0.807–1.211)	(0.947–1.534)	(1.569–1.765)	(1.416–1.536)
			P < 0.001	P = 0.004	P < 0.001	P = 0.913	P = 0.128	P < 0.001	P < 0.001
VTDR ^b	26,950	1,042 (3.9%)	0.901	1.244	0.691	1.037	1.419	1.729	1.574
			(0.852–0.953)	(1.091–1.418)	(0.581–0.823)	(0.862–1.247)	(1.152–1.747)	(1.640–1.823)	(1.517–1.633)
			P < 0.001	P = 0.001	P < 0.001	P = 0.703	P < 0.001	P < 0.001	P < 0.001
Cataract	27,415	639 (2.3%)	1.704	0.845	1.617	1.399	3.746	1.080	1.608
			(1.568–1.852)	(0.709–1.006)	(1.273–2.055)	(1.053–1.859)	(2.948–4.760)	(1.006–1.158)	(1.547–1.671)
			P < 0.001	P = 0.059	P < 0.001	P = 0.021	P < 0.001	P = 0.033	P < 0.001

^aReference category: Hispanic (the most prevalent race/ethnicity in this cohort). ^bPre-specified primary prediction tasks. Bold font indicates $P < 0.05$.

Table 3 | Adjusted analysis for validation set B

Prediction target	Number of visits		Odds ratio (95% CI), P value						
	Total	Positive (%)	Age (per decade)	Male	White ^a	Black ^a	Asian/Pacific islander ^a	Years with diabetes (per five years)	DLS (external eye image)
HbA1c ≥7%	4,120	2,819 (68.4%)	0.866	1.161	0.812	0.832	1.526	1.417	2.245
			(0.803–0.933)	(0.997–1.352)	(0.607–1.086)	(0.639–1.083)	(0.814–2.859)	(1.318–1.524)	(2.054–2.453)
			P < 0.001	<i>P</i> = 0.054	<i>P</i> = 0.160	<i>P</i> = 0.172	<i>P</i> = 0.187	P < 0.001	P < 0.001
HbA1c ≥8%	4,120	1,991 (48.3%)	0.909	1.205	0.842	0.690	1.482	1.315	2.293
			(0.844–0.978)	(1.047–1.388)	(0.634–1.118)	(0.538–0.885)	(0.817–2.690)	(1.235–1.400)	(2.101–2.501)
			P = 0.010	P = 0.009	<i>P</i> = 0.234	P = 0.003	<i>P</i> = 0.195	P < 0.001	P < 0.001
HbA1c ≥9% ^c	4,120	1,455 (35.3%)	0.945	1.175	0.834	0.788	1.418	1.197	2.241
			(0.875–1.021)	(1.017–1.358)	(0.615–1.130)	(0.609–1.020)	(0.754–2.666)	(1.124–1.274)	(2.052–2.448)
			<i>P</i> = 0.150	P = 0.029	<i>P</i> = 0.241	<i>P</i> = 0.070	<i>P</i> = 0.279	P < 0.001	P < 0.001
Tch ≥ 200 mg dl ⁻¹	859	219 (25.5%)	1.077	0.865	1.914	1.365	3.062	0.920	1.419
			(0.924–1.256)	(0.625–1.199)	(1.168–3.135)	(0.927–2.011)	(1.200–7.814)	(0.806–1.050)	(1.179–1.707)
			<i>P</i> = 0.343	<i>P</i> = 0.385	P = 0.010	<i>P</i> = 0.115	P = 0.019	<i>P</i> = 0.214	P < 0.001
Tch ≥ 240 mg dl ⁻¹	859	82 (9.5%)	1.074	0.788	1.709	1.318	0.675	0.821	1.316
			(0.860–1.342)	(0.492–1.262)	(0.855–3.418)	(0.752–2.312)	(0.087–5.273)	(0.671–1.005)	(1.032–1.679)
			<i>P</i> = 0.526	<i>P</i> = 0.322	<i>P</i> = 0.129	<i>P</i> = 0.335	<i>P</i> = 0.708	<i>P</i> = 0.057	P = 0.027
Tg ≥ 150 mg dl ⁻¹	843	414 (49.1%)	1.123	1.104	1.144	0.819	0.702	0.881	1.458
			(0.988–1.276)	(0.834–1.460)	(0.709–1.846)	(0.530–1.266)	(0.288–1.715)	(0.784–0.991)	(1.207–1.761)
			<i>P</i> = 0.075	<i>P</i> = 0.491	<i>P</i> = 0.582	<i>P</i> = 0.369	<i>P</i> = 0.438	P = 0.035	P < 0.001
Tg ≥ 200 mg dl ⁻¹	843	260 (30.8%)	1.072	1.153	1.289	0.625	0.712	0.923	1.404
			(0.928–1.238)	(0.849–1.566)	(0.791–2.099)	(0.386–1.011)	(0.253–2.008)	(0.812–1.048)	(1.156–1.705)
			<i>P</i> = 0.345	<i>P</i> = 0.362	<i>P</i> = 0.309	<i>P</i> = 0.056	<i>P</i> = 0.521	<i>P</i> = 0.215	P < 0.001
Tg ≥ 500 mg dl ⁻¹	843	21 (2.5%) ^b	0.770	2.492	3.288	0.500	2.044	0.789	0.890
			(0.479–1.238)	(0.930–6.677)	(0.972–11.121)	(0.107–2.344)	(0.237–17.615)	(0.525–1.183)	(0.527–1.504)
			<i>P</i> = 0.280	<i>P</i> = 0.069	<i>P</i> = 0.056	<i>P</i> = 0.379	<i>P</i> = 0.515	<i>P</i> = 0.251	<i>P</i> = 0.663
Mild+ DR	4,982	1,509 (30.3%)	0.937	1.358	0.784	0.763	1.883	1.710	2.897
			(0.873–1.006)	(1.169–1.578)	(0.589–1.042)	(0.586–0.993)	(1.017–3.488)	(1.604–1.822)	(2.667–3.147)
			<i>P</i> = 0.071	P < 0.001	<i>P</i> = 0.094	P = 0.044	P = 0.044	P < 0.001	P < 0.001
Moderate+DR ^c	4,982	1,172 (23.5%)	0.903	1.443	0.657	0.723	1.600	1.768	3.102
			(0.835–0.977)	(1.221–1.704)	(0.470–0.918)	(0.540–0.970)	(0.806–3.174)	(1.650–1.894)	(2.845–3.383)
			P = 0.011	P < 0.001	P = 0.014	P = 0.030	<i>P</i> = 0.179	P < 0.001	P < 0.001
Severe+ DR	4,982	455 (9.1%)	0.852	1.417	0.483	0.523	0.173	1.622	2.504
			(0.761–0.954)	(1.123–1.788)	(0.273–0.853)	(0.314–0.870)	(0.024–1.236)	(1.476–1.781)	(2.286–2.744)
			P = 0.006	P = 0.003	P = 0.012	P = 0.013	<i>P</i> = 0.080	P < 0.001	P < 0.001
DME ^c	4,982	394 (7.9%)	1.044	1.637	0.315	0.511	0.645	1.735	1.957
			(0.930–1.172)	(1.296–2.068)	(0.165–0.600)	(0.318–0.823)	(0.197–2.119)	(1.580–1.906)	(1.801–2.127)
			<i>P</i> = 0.466	P < 0.001	P < 0.001	P = 0.006	<i>P</i> = 0.470	P < 0.001	P < 0.001
VTDR ^c	4,982	617 (12.4%)	0.925	1.529	0.439	0.549	0.590	1.720	2.581
			(0.837–1.022)	(1.244–1.879)	(0.268–0.718)	(0.366–0.822)	(0.187–1.865)	(1.584–1.869)	(2.364–2.818)
			<i>P</i> = 0.125	P < 0.001	P = 0.001	P = 0.004	<i>P</i> = 0.369	P < 0.001	p < 0.001
Cataract	5,058	268 (5.3%)	1.465	1.320	1.371	0.647	3.364	1.259	2.695
			(1.258–1.706)	(0.961–1.815)	(0.791–2.374)	(0.348–1.200)	(1.509–7.500)	(1.116–1.419)	(2.451–2.964)
			P < 0.001	<i>P</i> = 0.087	<i>P</i> = 0.260	<i>P</i> = 0.167	P = 0.003	P < 0.001	P < 0.001

^aReference category: Hispanic (the most prevalent race/ethnicity in this cohort). ^bFewer than 50 positive examples. ^cPre-specified primary prediction tasks. Bold font indicates *P* < 0.05.

Table 4 | Adjusted analysis for validation set C

Prediction target	Number of visits		Odds ratio (95% CI), <i>P</i> value		
	Total	Positive (%)	Age (per decade)	Male	DLS (external eye image)
HbA1c ≥ 7%	8,995	5,119 (56.9%)	0.881 (0.841–0.923) <i>P</i> < 0.001	1.181 (0.958–1.456) <i>P</i> = 0.119	1.616 (1.541–1.695) <i>P</i> < 0.001
HbA1c ≥ 8%	8,995	2,947 (32.8%)	0.830 (0.788–0.875) <i>P</i> < 0.001	1.185 (0.953–1.474) <i>P</i> = 0.126	1.706 (1.618–1.798) <i>P</i> < 0.001
HbA1c ≥ 9% ^a	8,995	1,766 (19.6%)	0.807 (0.757–0.860) <i>P</i> < 0.001	1.004 (0.788–1.279) <i>P</i> = 0.973	1.770 (1.665–1.881) <i>P</i> < 0.001
Tch ≥ 200 mg dl ⁻¹	9,304	1,935 (20.8%)	0.805 (0.755–0.860) <i>P</i> < 0.001	0.802 (0.646–0.997) <i>P</i> = 0.047	1.255 (1.174–1.341) <i>P</i> < 0.001
Tch ≥ 240 mg dl ⁻¹	9,304	580 (6.2%)	0.768 (0.693–0.850) <i>P</i> < 0.001	0.813 (0.585–1.130) <i>P</i> = 0.217	1.272 (1.153–1.403) <i>P</i> < 0.001
Tg ≥ 150 mg dl ⁻¹	9,280	3,825 (41.2%)	0.943 (0.902–0.985) <i>P</i> = 0.009	1.533 (1.227–1.914) <i>P</i> < 0.001	1.745 (1.668–1.825) <i>P</i> < 0.001
Tg ≥ 200 mg dl ⁻¹	9,280	2,261 (24.4%)	0.908 (0.863–0.956) <i>P</i> < 0.001	1.870 (1.409–2.481) <i>P</i> < 0.001	1.759 (1.673–1.849) <i>P</i> < 0.001
Tg ≥ 500 mg dl ⁻¹	9,280	205 (2.2%)	0.791 (0.678–0.922) <i>P</i> = 0.003	1.368 (0.630–2.973) <i>P</i> = 0.428	1.579 (1.419–1.757) <i>P</i> < 0.001
Mild+DR	9,518	992 (10.4%)	0.945 (0.881–1.013) <i>P</i> = 0.112	1.614 (1.078–2.417) <i>P</i> = 0.020	2.000 (1.892–2.114) <i>P</i> < 0.001
Moderate+DR ^a	9,518	454 (4.8%)	0.837 (0.756–0.925) <i>P</i> < 0.001	1.768 (0.968–3.229) <i>P</i> = 0.064	2.044 (1.913–2.184) <i>P</i> < 0.001
Severe+DR	9,518	203 (2.1%)	0.795 (0.689–0.917) <i>P</i> = 0.002	2.571 (0.937–7.057) <i>P</i> = 0.067	1.640 (1.525–1.763) <i>P</i> < 0.001
DME ^a	9,528	119 (1.2%)	0.590 (0.491–0.709) <i>P</i> < 0.001	2.135 (0.729–6.252) <i>P</i> = 0.167	1.746 (1.583–1.926) <i>P</i> < 0.001
VTDR ^a	9,528	287 (3.0%)	0.725 (0.640–0.822) <i>P</i> < 0.001	2.043 (0.948–4.405) <i>P</i> = 0.068	1.962 (1.828–2.106) <i>P</i> < 0.001

^aPre-specified primary prediction tasks. Bold font indicates *P* < 0.05.

DLS did not achieve higher AUC compared with the baseline model, the adjusted odds ratios were statistically significant (*P* < 0.05) in all cases, except where there were fewer than 50 positives (triglycerides ≥ 500 mg dl⁻¹, the only example among 20 analyses).

Explainability analysis. To better understand how the DLS was able to detect these diseases without looking inside the eye at the retina, we next conducted several explainability experiments. Predicting

cataract presence acted as a positive control because it manifests as media opacity ('cloudiness' of the lens) and, thus, is generally visible from external eye images. The AUC of the DLS for predicting cataract was substantially higher than that for predicting the other targets, at 86.4% and 93.2% for validation sets A and B, respectively (labels were not available for sets C and D) (Supplementary Table 1).

In ablation analysis, we removed portions of each image, trained another DLS on such images and then evaluated this DLS

Table 5 | Adjusted analysis for validation set D

Prediction target	Number of visits		Odds ratio (95% CI), <i>P</i> value					
	Total	Positive (%)	Age (per decade)	BMI (per 5 kg m ⁻¹)	Male	White ^a	Other ^a	DLS (external eye image)
HbA1c ≥ 7%	2,949	1,045 (35.4%)	1.440 (1.328–1.562) <i>P</i> < 0.001	1.413 (1.320–1.512) <i>P</i> < 0.001	1.574 (1.196–2.071) <i>P</i> = 0.001	0.790 (0.659–0.946) <i>P</i> = 0.010	0.900 (0.626–1.295) <i>P</i> = 0.571	1.896 (1.735–2.073) <i>P</i> < 0.001
HbA1c ≥ 8%	2,949	572 (19.4%)	1.315 (1.191–1.451) <i>P</i> < 0.001	1.230 (1.141–1.325) <i>P</i> < 0.001	1.651 (1.183–2.305) <i>P</i> = 0.003	0.656 (0.529–0.814) <i>P</i> < 0.001	0.692 (0.442–1.083) <i>P</i> = 0.107	1.966 (1.772–2.182) <i>P</i> < 0.001
HbA1c ≥ 9% ^b	2,949	331 (11.2%)	1.235 (1.088–1.401) <i>P</i> = 0.001	1.170 (1.068–1.283) <i>P</i> < 0.001	1.707 (1.147–2.542) <i>P</i> = 0.008	0.387 (0.290–0.516) <i>P</i> < 0.001	0.341 (0.173–0.672) <i>P</i> = 0.002	1.992 (1.759–2.256) <i>P</i> < 0.001
Tch ≥ 200 mg dl ⁻¹	3,866	1,029 (26.6%)	0.917 (0.849–0.992) <i>P</i> = 0.030	0.871 (0.819–0.926) <i>P</i> < 0.001	0.624 (0.509–0.764) <i>P</i> < 0.001	1.043 (0.888–1.226) <i>P</i> = 0.606	1.171 (0.853–1.606) <i>P</i> = 0.329	1.280 (1.168–1.402) <i>P</i> < 0.001
Tch ≥ 240 mg dl ⁻¹	3,866	301 (7.8%)	0.948 (0.841–1.069) <i>P</i> = 0.382	0.809 (0.728–0.899) <i>P</i> < 0.001	0.562 (0.414–0.762) <i>P</i> < 0.001	1.171 (0.899–1.524) <i>P</i> = 0.242	1.431 (0.883–2.320) <i>P</i> = 0.146	1.213 (1.062–1.385) <i>P</i> = 0.004
Tg ≥ 150 mg dl ⁻¹	3,891	1,102 (28.3%)	1.005 (0.935–1.081) <i>P</i> = 0.884	1.164 (1.096–1.236) <i>P</i> < 0.001	1.778 (1.381–2.290) <i>P</i> < 0.001	1.300 (1.033–1.637) <i>P</i> = 0.026	1.330 (0.954–1.854) <i>P</i> = 0.092	1.552 (1.394–1.728) <i>P</i> < 0.001
Tg ≥ 200 mg dl ⁻¹	3,891	612 (15.7%)	0.998 (0.912–1.093) <i>P</i> = 0.967	1.178 (1.096–1.266) <i>P</i> < 0.001	1.507 (1.094–2.075) <i>P</i> = 0.012	1.376 (1.043–1.815) <i>P</i> = 0.024	1.304 (0.869–1.958) <i>P</i> = 0.200	1.533 (1.360–1.728) <i>P</i> < 0.001
Tg ≥ 500 mg dl ⁻¹	3,891	43 (1.1%) ^c	0.828 (0.607–1.128) <i>P</i> = 0.231	1.044 (0.821–1.329) <i>P</i> = 0.724	3.271 (0.752–14.235) <i>P</i> = 0.114	1.652 (0.753–3.625) <i>P</i> = 0.211	0.999 (0.221–4.518) <i>P</i> = 0.999	1.213 (0.936–1.571) <i>P</i> = 0.144

^aReference category: Black (the most prevalent race/ethnicity in this cohort). ^bPre-specified primary prediction tasks. ^cFewer than 50 positive examples. Bold font indicates *P* < 0.05.

(Methods)³. The central region of the image was more important than the outer rim for all predictions except for lipids, with the biggest delta being our positive control, cataract (Fig. 2a,b). For lipids predictions, the central region and outer rim appeared equally important.

Next, we conducted saliency analysis using several gradient-based methods (Methods and Fig. 3). The systemic predictions (HbA1c and lipids) were the most distinct, with almost all of the attention on the nasal and temporal conjunctiva/sclera (that is, where conjunctival vessels tend to be prominent; Fig. 3b–d). By contrast, all three diabetic retinal disease predictions (moderate+ DR, DME and VTDR) focused on both the pupil and the nasal and temporal iris and conjunctiva (Fig. 3e–g). The cataract predictions focused on the pupil in the middle of the image (Fig. 3h). The results of additional saliency methods are presented in Extended Data Fig. 4a–g, with a quantitative summary of these trends in Extended Data Fig. 4h,i.

Because diabetes is known to affect pupil sizes^{7,12}, and multiple attribution methods highlighted the pupil/iris region (Fig. 3 and Extended Data Fig. 4), we reasoned that the DLS could have learnt to perform these predictions via measuring pupil size. To test this hypothesis, we first trained a pupil/iris detection model using a subset of the development set (Methods), and we used this model to estimate the pupil size in the validation sets (Supplementary Information and Supplementary Figs. 1 and 2). To correct for differences in distance from the camera, we normalized the pupil size by representing the pupil radius as a fraction of the iris radius. The iris size is equivalent to the corneal diameter since the visible iris

extends to the edge of the corneoscleral junction. Corneal diameter (or iris size here) is known to be relatively constant in the adult population^{13,14}.

We first examined to what degree pupil size added useful information to these predictive tasks by inspecting the effect of its addition to the baseline characteristics model (Supplementary Table 2). There was generally little to no improvement in AUC of the baseline characteristics model (approximately 1% with overlapping confidence intervals (CIs)). The only exceptions were in validation set C, with 3–4% AUC improvements for moderate+ DR and VTDR. The improvement of the DLS over these augmented baseline characteristics models remained statistically significant in all cases (*P* < 0.001). Similarly, in an adjusted analysis including estimated pupil size, the external eye predictions retained statistically significant odds ratios (Supplementary Table 3).

Additionally, to determine whether the model's ability to detect diabetic retinal disease could be explained by its ability to detect elevated HbA1c, we conducted an analysis adjusting for laboratory-measured HbA1c values (Supplementary Table 4). Even in this adjusted analysis, the odds ratios remained statistically significant, ranging from 1.4 to 3.2 on the primary predictions (*P* < 0.001 for moderate+ DR, DME and VTDR across all three validation datasets that contain DR labels).

Sensitivity and subgroup analysis. Because these external eye images were taken with fundus cameras, we next sought to examine whether lower-quality images would suffice. To simulate

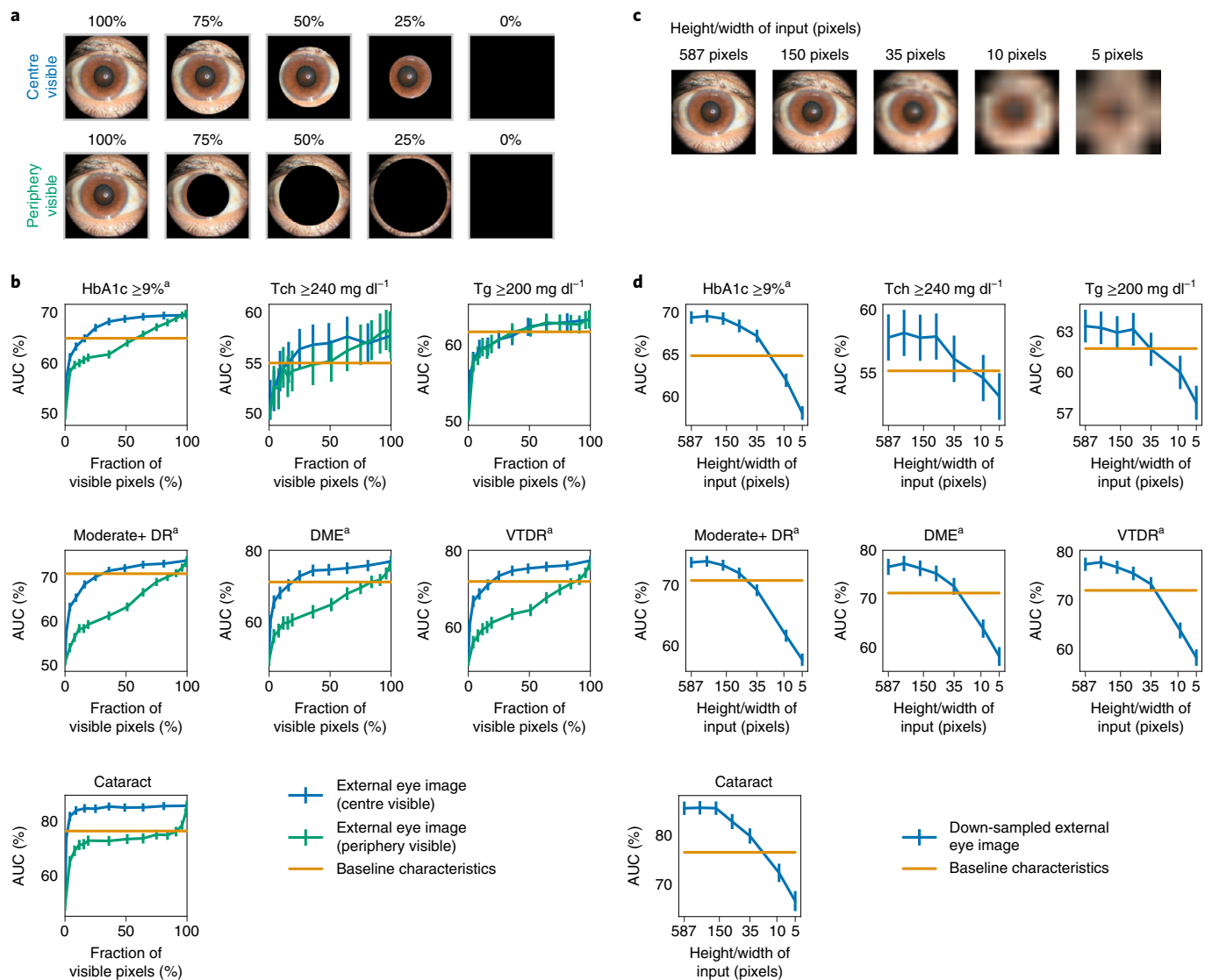


Fig. 2 | Importance of different regions of the image and impact of image resolution. a, Sample images with different parts of the image visible. Top: only a circle centred around the centre is visible. Bottom: only the peripheral rim is visible. **b**, AUCs as a function of the fraction of input pixels visible when the DLS sees only the centre (blue curve) or peripheral regions (green curve). A large delta between the blue and green curves (for example, for cataract) indicates that most of the information is in the region corresponding to the blue line. **c**, Sample images of different resolutions. Images were down-sampled to a certain size (for example, '5' indicates 5×5 pixels across) and then up-sampled to the same original input size of 587×587 pixels to ensure that the network has the exact same capacity (Methods). **d**, AUCs corresponding to different input image resolutions. Error bars are 95% CIs computed using the DeLong method. ^aPre-specified primary prediction tasks.

low-quality images, the input images were down-sampled to lower resolutions for both training and evaluation (Methods). The DLS performance decreased slightly as the input size decreased to 75 pixels and then substantially as the input size decreased below 35 pixels (Fig. 2c,d).

Next, we conducted extensive subgroup analyses across multiple variables: demographic information (age, sex and race/ethnicity), presence of cataract, camera types (for a subset of data where this information was available) and pupil size in validation set A (Supplementary Tables 5–11). In brief, the DLS (and the baseline variables) remained predictive in all subgroups. The DLS consistently outperformed the baseline characteristics model, albeit with the difference not reaching statistical significance in some subgroups. The DLS AUC dropped by more than 5% in some subgroups: the triglycerides ≥ 200 and moderate+ DR targets in the Black subgroup; the VTDR target in the >10 years with diabetes

subgroup; for cataract in the age >50 years, small pupil and camera type B subgroups. In all of these cases, the AUC of the baseline characteristics model decreased as well.

Validation set D had additional variables available: body mass index (BMI), diabetes status and presence of intraocular lens (IOL). We conducted subgroup analysis based on these variables for both the HbA1c $\geq 9\%$ and $\geq 7\%$ prediction tasks (due to the relatively few numbers of positive examples of the former; Supplementary Tables 12 and 13). Across the different BMI groups, the DLS AUC dropped by at most 4% compared with the full dataset. In the non-diabetic group, as may be expected, there were few positives for these cut-offs and, hence, CIs are wide. Nevertheless, the DLS outperformed the baseline characteristics model (57.6% versus 53.7% for HbA1c $\geq 7\%$). Across the two IOL groups, AUC differed by at most 0.6% compared with the full dataset. Taking into account both HbA1c $\geq 9\%$ and HbA1c $\geq 7\%$, we did not observe any trends

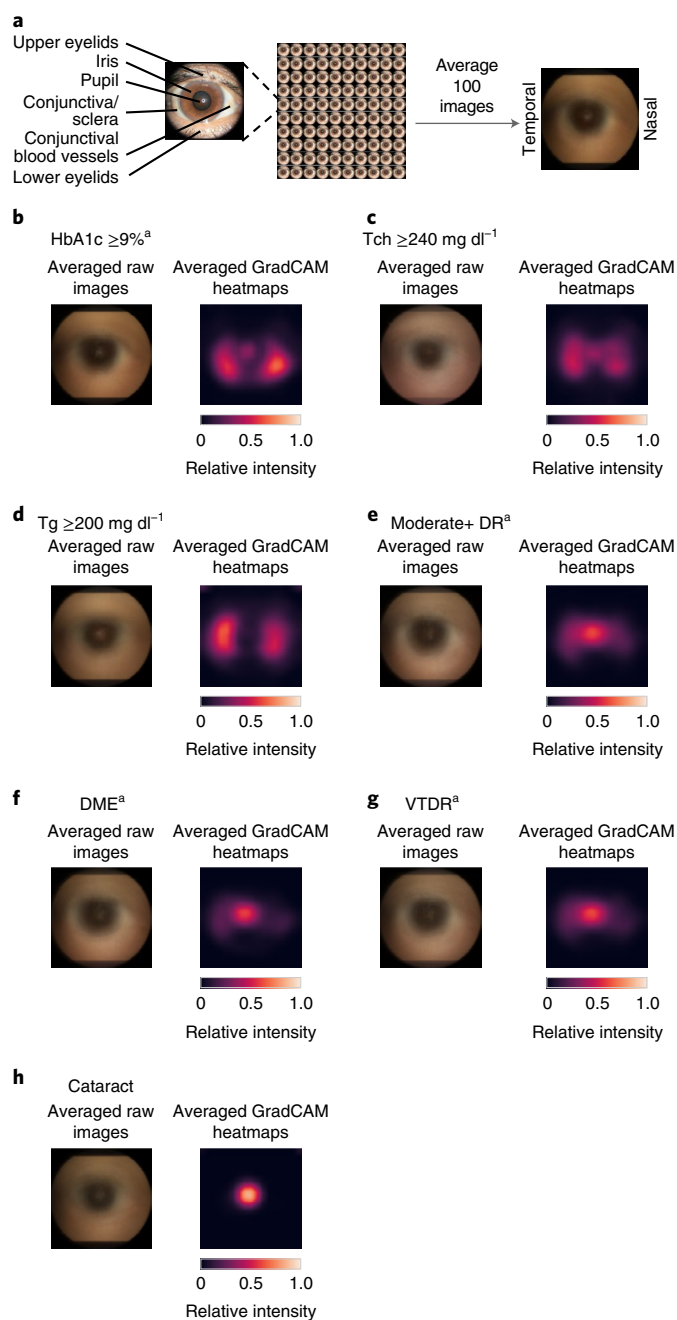


Fig. 3 | Qualitative and quantitative saliency analysis illustrating the influence of various regions of the image towards the prediction.

a, Overview of anterior eye anatomy. **b–h**, Eye images from patients with HbA1c $\geq 9\%$ (**b**), Tch ≥ 240 mg dl⁻¹ (**c**), Tg ≥ 200 mg dl⁻¹ (**d**), moderate+ DR (**e**), DME (**f**), VTDR (**g**), and cataract (**h**). Illustrations are the result of averaging images or saliency maps from 100 eyes, with laterality manually adjusted via horizontal flips to have the nasal side on the right (Methods). Left: the original averaged image. Right: a similarly averaged saliency from GradCAM. Saliency maps from other methods are shown in Supplementary Fig. 4a–g. ^aPre-specified primary prediction tasks.

towards worse performance with race/ethnicity in this dataset with a higher proportion of Black patients.

Discussion

Our results show that external images of the eye contain signals of diabetes-related retinal conditions, poor blood sugar control and

elevated lipids. We accomplished this through the development of a DLS, which generalized to diverse patient populations (including race/ethnicity groups with differences in iris colour and the red reflex), different imaging protocols and devices from independent clinics in multiple US states. The results for our primary prediction targets (HbA1c $\geq 9\%$ and presence of varying levels of diabetic retinal disease) were statistically significantly better compared with using demographic information and medical history (such as years with diabetes) alone and remained statistically significant after adjusting for multiple baseline characteristics and within numerous subgroups. Our exploratory results also show promise in extending this method to predicting elevated lipids and generalizing to a non-diabetic population.

The discovery that predictions about systemic parameters and diabetic retinal disease could be derived from external eye photography is surprising, since such images are primarily used to identify and monitor anterior eye conditions, such as eyelid and conjunctival malignancies, corneal infections and cataracts. Although there are numerous studies noting that conjunctival vessel changes (fewer, wider and less tortuous conjunctival vessels)^{9,10,15–17} are associated with duration of diabetes^{18,19} and severity of DR^{20,21}, and elevated cholesterol levels and atherosclerosis have been linked to xanthelasma (yellowish deposit under the eyelid skin)²². However, to our knowledge, there have been no large studies linking HbA1c or diabetic macular oedema to conjunctival vessel changes in diabetes. Furthermore, conjunctival vessel assessment for signs of diabetes or elevated lipids is not a common clinical practice due to the subjectivity and time-consuming nature of such an evaluation and the option of a more accurate and easier test for the clinician (HbA1c). We verified that these surprising results were reproducible and not an artefact of a single dataset or site via broad geographical validation across 18 US states.

Our evidence provides several hints as to how these predictions are possible. First, the ablation analysis indicates the centre of the image (pupil/lens, iris/cornea and conjunctiva/sclera) is substantially more important than the image periphery (for example, eyelids) for all predictions (Fig. 2a,b). Second, the saliency analysis similarly indicates that the DLS is most influenced by areas near the centre of the image. These included the pupil and the corneo-scleral junction for diabetic retinal disease and both the nasal and temporal conjunctiva (where conjunctival blood vessels are often most prominent) for blood sugar control (HbA1c) (Fig. 3). In both analyses, the positive control (cataract, which is visible at the pupil/lens) provides a useful baseline attribution map for a prediction that is expected to focus exclusively on the pupil/lens. Moreover, the quantitative results where the DLS remained predictive within pupil size subgroups, after adjusting for pupil size, and compared with a baseline model augmented with pupil size, suggest that the performance of the DLS could not be explained by it relying on pupil size alone. Together, these data suggest that the DLS is leveraging both information in the pupil region, such as from the light reflecting from the retina ('red reflex'), and information outside of the pupil/iris region, such as in the conjunctival vessels. For diabetic retinal disease predictions, we observed that DLS scores remained statistically significant even when adjusting for HbA1c, suggesting that the model is picking up on signals beyond elevated HbA1c to make these predictions. Finally, the performance of the DLS for HbA1c trended upwards for higher HbA1c cut-offs (AUC: 66.9% for $\geq 7\%$ versus 69.1% for $\geq 8\%$ versus 70.0% for $\geq 9\%$ in validation set A), compared with a 'flat' trend of 65% for baseline characteristics. This 'dose dependent' trend cannot be explained by increased training data (because the number of examples of $\geq 9\%$ is strictly less than $\geq 8\%$) and suggests that the features used by the DLS were related to the extent of elevated glucose levels. A similar trend of increasing DLS AUCs for higher cut-offs was observed for total cholesterol and triglycerides. Better scientific understanding of these predictions

via systematic falsifying of hypotheses, and potentially enabling experts to perform the same predictions, will need to be addressed in future work.

Our technique has potential implications for the large and rapidly growing population of patients with diabetes because it does not, in principle, require specialized equipment. Specifically, screening for diabetes-related retinal disease requires funduscopy or the use of a fundus camera to examine the back of the eye through the pupil. This limits disease screening and detection examinations to either eye clinics or store-and-forward teleretinal screening sites where fundus cameras are present—both of which require in-person visits. Indeed, poor access and cost contribute, in part, to low DR screening rates²³. Similarly, HbA1c and lipid measurements require an in-person visit for a venous blood draw, which can be unpleasant for patients and have multiple potential side effects, including bleeding, bruising and nerve damage²⁴. Our results suggest that signs of these conditions are extractable from a photograph of the front of the eye without pupil dilation via eye drops, as shown in validation set A. There appeared to be performance differences in the proposed model with versus without dilation, but the prevalence differences and the fact that baseline models (that do not depend on whether the eyes were dilated) shifted in the same direction suggest that a patient population difference was at least partially responsible. While the experiments in the paper did use external eye images captured by a fundus camera, we show that even low-resolution images of 75 × 75 pixels (which is 1% of the resolution of a basic ‘720p’ laptop webcam and 0.1% of the resolution of a standard 8-megapixel smartphone camera) result in adequate performance, suggesting that the resolution requirements for this technique can be easily met. While further work is needed to determine whether there are additional requirements for lighting, photography distance or angle, image stabilization, lens quality or sensor fidelity, we hope that disease detection techniques via external eye images can eventually be widely accessible to patients, whether in clinics, pharmacies or even at home. The AUCs in our experiments may not be adequate for a confirmatory diagnosis; for example, it is not expected to approach detection parity with the reference standard obtained via fundus photographs or blood test. However, they are comparable to other low-cost and accessible screening tools, such as the pre-diabetes risk score from the US Centers for Disease Control and Prevention^{25,26}, which achieves an AUC of 0.79–0.83, with a PPV ranging from 5% to 13%. Because lack of access to and underuse of healthcare are associated with poor diabetic management²⁷, we hope that the availability of easily accessible techniques such as ours can improve adherence to diabetic interventions, which we know to be cost saving or cost effective²⁸. Finally, our exploratory analysis on validation set D shows promise in extending this approach to a more general, non-diabetic population.

The potential future use cases for easy identification and monitoring of high-risk diabetic patients are manifold. First, detecting patients with diabetes who have difficulty controlling their blood glucose (HbA1c ≥ 9%) may help reveal which patients are in need of further counselling, additional diabetic resources and medication changes. In our analysis, when the top 5% of patients with the highest predicted likelihood were examined, 50–75% may truly have HbA1c ≥ 9%. Similarly, our exploratory analysis on the non-diabetic subgroup of validation set D demonstrates that it may be possible to identify potentially asymptomatic patients at risk for diabetes and help determine which patients may benefit from a confirmatory blood test and early interventions, such as lifestyle counselling or medications. Second, identification of patients at risk for diabetic retinal disease can determine which patients may benefit from ophthalmology follow-up and targeted treatment to avoid diabetes-associated vision loss. In our analysis, if the top 5% of patients with the highest predicted likelihood of various severities of diabetic retinal disease were examined via fundus photographs, 20–65% could have vision-threatening diabetic retinal disease, and

20–90% could have moderate-or-worse diabetic retinal disease that warrants ophthalmology referral. Importantly, compliance with diabetic eye disease screening is estimated to be only 20–50% (refs. 29–32) due to factors such as transportation difficulties²³. If our approach could be shown to generalize to images captured by non-specialized equipment, remote DR ‘pre-screening’ using this method may enable targeted support for high-risk patients. Conversely, patients at very low risk of developing diabetic retinal disease could potentially be screened less frequently. Additionally, detection of high blood lipids, or hyperlipidemia, from external eye photos allows for an easy, non-invasive approach for cardiovascular disease risk factor³³ screening. Early identification of individuals with hyperlipidemia, who have a twofold increased risk of cardiovascular disease development³⁴, could allow for earlier lifestyle counselling, medication intervention, additional disease screening and reduction in morbidity and mortality. Lastly, minimizing in-person visits to health clinics has become exceedingly important in preventing disease spread. During the current coronavirus disease 2019 (COVID-19) pandemic, many countries are actively encouraging ‘social’ (physical) distancing and deferment of non-urgent medical visits, including decreased visits for diabetes³⁵. Similarly to how phone cameras^{36,37} can help identify digital biomarkers of diabetes³⁸, external eye photo-based screening could allow for remote evaluation, potentially from home, and with common equipment, greatly minimizing risk to the individual and others.

This work establishes that signs of systemic disease can be identified from external eye images. These initial findings raise the tantalizing prospect that such external eye photographs may contain additional useful signals, both familiar and novel, related to other systemic conditions. Clinically, numerous external or anterior segment findings are correlated with or are the result of underlying systemic disease. Obstructive sleep apnoea is associated with floppy eyelid syndrome (easily everted eyelids) and resultant papillary conjunctivitis^{39,40}. Thyroid disease can manifest specific ocular signs such as upper eyelid retraction, conjunctival injection (redness) and chemosis (swelling) and caruncular oedema (swelling of small, pink nodule at the inner corner of the eye)^{41,42}. Atherosclerosis has been linked with xanthelasma, which is predictive of adverse cardiac outcomes⁴³. Systemic hypertension, another high-morbidity disease, has been associated with conjunctival microangiopathy and correlates with time since hypertension diagnosis⁴⁴. Beyond eyelid and conjunctival changes, deposition of calcium or uric acid in the cornea may signify derangements related to hyperparathyroidism, chronic renal failure and gout^{45–47}. Previous work has also shown that the severity of corneal and conjunctival calcium deposition, as determined by external eye photos, is predictive of all-cause one-year mortality risk in haemodialysis patients⁴⁸. Recently, hepatobiliary disease signals have been identified in slit-lamp photos of the eye⁴⁹. These manifestations could be readily captured with external eye photography that documents the eyelids, conjunctiva and cornea, and further work examining these disease populations is needed to assess additional systemic disease prediction models. Similarly to how investigations into manifestations of systemic disease in the retina have been dubbed ‘oculomics’⁵⁰, such analyses on external ocular images could be termed ‘exoculomics’.

Our study is not without limitations. First, we have limited data with which to understand if smartphone or webcam images are sufficient, because all images were taken by healthcare professionals using a table-top fundus camera. Although the resolution and actual image sensor (for example, Canon EOS-series cameras) were similar to consumer-grade cameras, these clinical-grade cameras provided chin stabilization that may have reduced motion blur and controlled the shooting angle and distance, good lighting via circular flash to reduce occlusion of features from shadows and greater magnification⁵¹. Similarly, despite only semi-standardized protocols, clinical professionals may have provided instructions to standardize gaze

and improve image quality, and the clinical environment may have been darkened to enlarge pupil size. Additional work is needed to evaluate whether the DLS generalizes to other consumer-grade cameras and settings without modification or if additional data collection for further training is needed. Second, part of the model's performance may be derived from detecting confounding factors not available in this study. For example, patients with diabetes are more likely to develop cataracts⁵² and other anterior segment signs of proliferative DR (such as rubeosis iridis⁵³ or neovascular glaucoma⁵⁴) than non-diabetics, although these conditions are not common and unlikely to explain all of the model's performance. To control confounding effects from cataract, we showed that the image-based model performed better than the baseline even without the central, pupil-containing region, and we further verified our findings in a subgroup excluding patients labelled with cataract presence. However, cataract was not the primary screening goal in these datasets, and, therefore, their recall was probably imperfect; cataract type and severity were also unavailable. Other potentially relevant variables, such as BMI⁵⁵ or use of related medications, were not consistently available and, thus, could not be used to develop baseline models or perform adjustment. Further study with verified and more granular incidental findings will help quantify the contribution from such potential confounding factors. Additionally, while we did present exploratory analysis showing promise in generalizing to non-diabetic populations, this will need to be evaluated further on a larger dataset, and the model may benefit from further training or refinement on this specific patient population (that is, those without diabetes). Lastly, other imaging modalities, such as slit-lamp photography, can reveal further signs of non-ophthalmic issues (hepatobiliary disease) from the eye⁴⁹, and future work in understanding how to integrate these modalities or extracting similar information from external eye photographs is needed.

In addition to the above limitations, as this is an initial study examining the feasibility of extracting useful signals of disease from external eye photographs, future work will be needed to tackle additional challenges before clinical validation studies. First, use cases and target patient populations should be identified, and the models should be thoroughly tested to ensure adequate accuracy across subgroups. If needed, the model should be further refined or tuned. Similarly, operating points should be customized for the target patient population and use case. Finally, whether the availability of this technology would improve adherence to diabetic retinal disease screening and ultimately improve patient outcomes will also need to be studied⁵⁶.

In summary, we demonstrated that external eye images can be used to detect the presence of several conditions, such as poor blood sugar control, elevated lipids and various diabetic retinal diseases. Further study is warranted to evaluate whether such a tool can be used in a home, pharmacy or primary care setting to improve disease screening and help with management of diabetes.

Methods

Datasets. This work used three teleretinal diabetic screening datasets and one more general eye care dataset in the United States (Table 1). The first two datasets were the California and non-California cohorts from EyePACS, a teleretinal screening service in the United States⁵⁷. In the EyePACS datasets, patients presented to sites, such as primary care clinics, for diabetic retinal disease screening. Visits from unknown sites were excluded, and patients associated only with sites from unknown states were excluded. We leveraged data from California (the state with the most visits in EyePACS) for development and the remaining US states for validation purposes. Within the development dataset, data from Los Angeles County were further excluded as held-out data for another project. Validation set A consisted of all non-California visits without pupil dilation, and validation set B consisted of all non-California visits with pupil dilation. The third validation dataset (C) was from the Atlanta VA Healthcare system's diabetic teleretinal screening programme⁵⁸, which serves multiple community-based outpatient clinics in the greater Atlanta area. Most visits in validation set C involved pupil dilation. The final validation dataset (D) was from the TECS programme established in primary care clinics in Georgia associated with the Atlanta VA Healthcare

system's catchment⁵⁹. Ethics review and institutional review board exemption for this retrospective study on de-identified data were obtained via the Advarra institutional review board.

Imaging protocol. As part of the standard imaging protocol for DR screening at these sites, patients had photographs taken of the external eye for evaluation of the anterior segment (along with the retinal fundus photograph). At sites served by EyePACS, cameras (where this information was available) included Canon (CR1 and CR2), Topcon (NW200 and NW400), Zeiss Visucam, Optovue iCam and Centervue DRS. On the basis of available image Exif metadata, images were digitized using Canon EOS single-lens reflex cameras. The image protocol involved first positioning patients comfortably in front of the camera with their chin resting on the chin rest and their forehead an inch away from the forehead brace. The operator then took external images of the right eye and then the left eye, ensuring image clarity, focus and visibility of the entire eye, including the iris, pupil and sclera⁶⁰. At sites served by the VA diabetic teleretinal screening programme, the cameras included Topcon NW8 and Topcon NW400. The acquisition protocol involved slightly distancing the patient from the fundus camera to capture a clear view of the external eye; further alignment was entrusted to the operator. The external eye photographs were intended to document findings such as cataracts and eyelid lesions. None of the external eye images was excluded for image quality or other reasons.

Labels. Diabetic retinal disease in EyePACS (development set and validation sets A and B) was graded by EyePACS-certified graders using a modified Early Treatment Diabetic Retinopathy Study grading protocol^{61,62}. Graders evaluated three-field retinal fundus photographs (nasal, primary and temporal), and the absence or presence of lesions was mapped⁶³ to five International Clinical Disease Severity Scale (ICDR) DR levels: no DR, mild, moderate, severe and proliferative DR and DME. VTDR was defined as severe-or-worse DR or DME. Diabetic retinal disease in the VA dataset (validation set C) was graded by ophthalmologists at the VA using the same protocol as EyePACS, except that only macular-centred and disc-centred fundus photographs were available. In all datasets, hard exudates were used as a surrogate for the presence of DME.

Baseline characteristics, including demographic information, were patient reported and provided by each site. Blood lab measurements (HbA1c, total cholesterol and triglycerides) were extracted from the patients' medical records and could be from prior visits. For the VA datasets, labs from more than 90 days away from the date of the external eye photo were discarded; for EyePACS, this was not possible due to lack of precise dates. Diabetes status for the VA TECS dataset was determined from both the International Classification of Diseases codes (ICD-10 codes E09–E11 and E13 and ICD-9 codes 250 and 362.0) available in their medical records and notes from their eye care visit (for example, if the patient self-reported as being diagnosed with DR, they were labelled as diabetic because the DR diagnosis requires a diabetes diagnosis). Intraocular lens status in the VA TECS dataset was determined from the notes of the eye care visit. Cataract presence in the EyePACS images was indicated by the EyePACS graders but was not available in the VA datasets.

Pupil size estimation. To better understand whether pupil size was associated with our results, we obtained segmentations of both pupil and iris size. To do so, 12 ophthalmologist graders evaluated a total of 5,000 external eye images. If the pupil and iris were distinct enough to delineate the borders accurately, the graders drew ellipses around both the iris and pupil (Supplementary Fig. 1). This was performed for a subset of 4,000 randomly selected images in the development set, another 500 images in validation sets A and B combined and 500 images in validation set C. We then trained a model to segment the iris and pupil using the labelled development set images and evaluated the accuracy in the validation sets (Supplementary Note 1 and Supplementary Fig. 2). Our pupil size analyses were based on running this model across all images.

DLS development. The DLS takes an external eye photograph as input and was trained to provide predictions about both systemic parameters (from labs: HbA1c, total cholesterol and triglycerides) and diabetic retinal disease status (graded via colour fundus photographs). The DLS does not 'see' the colour fundus photographs, whether during training or evaluation. To develop the external eye DLS, we split the development dataset into training and tuning, split in a 7:1 ratio, while ensuring that each site was only in one split. All visits for all patients were used for training or tuning. A multi-task learning approach was used to train a single network to predict all tasks (that is, one 'head' per task). Specifically, all heads adopted a classification setup (with cross-entropy loss) to ensure that all losses were in comparable units. For HbA1c and lipids, we opted for a classification instead of regression setup, because experiments in the tune set found a 1–2% performance improvement in AUC by adopting a classification approach relative to a regression approach (and using the predictions as a proxy for the predicted likelihood of each binary outcome). Specifically, HbA1c was trained as 3 binary classification heads with thresholds at 7%, 8% and 9%; total cholesterol was trained as two binary classification heads with thresholds at 200 mg dl⁻¹ and 240 mg dl⁻¹, and triglycerides were trained as three binary classification heads with thresholds

at 150 mg dl⁻¹, 200 mg dl⁻¹ and 500 mg dl⁻¹. DR was discretized into the five ICDR categories; DME was binary (present/absent); VTDR was binary (present/absent) and cataract was binary (present/absent). Some additional 'auxiliary' heads were included: age was discretized into seven categories: (0, 30], (30, 40], (40, 50], (50, 60], (70, 80] and (80, 90] years; race/ethnicity was categorized into Hispanic, White, Black, Asian and Pacific Islander, Native American and others; sex was represented as male and female; and years with diabetes was discretized into five categories: (0, 1.5], (1.5, 5.0], (5.0, 10.5], (10.5, 15.5] and (15.5, infinity) years. Because not every label was available for all of these heads, the losses were propagated only for the relevant heads for each example.

We trained five models based on the Inception-v3⁶⁴ architecture using hyperparameters summarized in Supplementary Table 14, and the predicted likelihood was averaged unless otherwise noted. Within the hyperparameter search space detailed in Supplementary Table 14, 100 random trials were conducted, and the top 5 models were selected. During each training run, the tune set AUC for HbA1c $\geq 7\%$ was the first to decline after reaching the peak, so early stopping was based on the tune set AUC of this task.

DLS evaluation. For evaluation, we selected a single visit per patient at random. This resulted in 27,415 visits for validation set A, 5,058 visits for set B and 10,402 visits for set C. We preprocessed the external eye photographs in the same manner as in previous work for colour fundus photographs¹⁶⁵. Some of the prediction targets (described below) were on a per-visit basis (for example, HbA1c has a single value per visit per patient), and so the predictions of both eyes were averaged for evaluation purposes. For prediction targets that were specifically for each eye (for example, eye diseases that affect each eye individually), we randomly selected one eye per visit during evaluation. The exact numbers of data points used for each prediction task are presented in Supplementary Table 1.

For validation sets A and B, visits without dilation status information were excluded (9% of the total visits). A small number of patients had both visits with dilation and visits without dilation and were included in both datasets ($n = 497$).

Baseline models for comparisons. Models leveraging baseline characteristics were used as a baseline and were trained using logistic regression with class-balanced weighting and the default L2 regularization (with the hyperparameter $C = 1.0$) in the scikit-learn Python library. For validation sets A and B, the input baseline characteristics were self-reported variables: age, sex, race/ethnicity and years with diabetes. These baseline models were trained using only undilated or dilated visits, respectively, because using the entire training dataset did not improve tune set performance. This also kept the development process of the baseline model consistent with that of the DLS. For validation sets C and D, because race/ethnicity and years with diabetes were not available, only age and sex were used as input variables. These baseline models were trained using the respective validation sets due to the large differences in patient population (mostly male and older and enriched for Black and White patients) and differences in available baseline characteristics. This overestimates the baseline model's performance and underestimates the improvement (delta) of our DLS.

Statistical analysis. We evaluated all results using the AUC, expressed as percentage (that is, 0–100%, where 50% is random performance). To evaluate the superiority of DLS predictions compared with the baseline model predictions in each dataset, we used the DeLong method⁶⁶. The superiority analysis on four tasks (HbA1c $\geq 9\%$, moderate+ DR, DME and VTDR) was pre-specified and documented as primary analyses before applying the DLS to the validation sets. Alpha was adjusted using the Bonferroni method for multiple comparison correction ($\alpha = 0.025$ for one-sided superiority test, divided by four tasks = 0.00625).

Pre-specified secondary analyses included additional evaluation metrics (sensitivity, specificity, PPV and NPV), additional related prediction tasks (cataract, HbA1c, total cholesterol, triglycerides, mild+ and severe+ DR), subgroup analysis (presence of cataract, pupil size, HbA1c stratification, demographic groups based on age, sex and race/ethnicity), adjusted analysis, explainability analysis and sensitivity analysis for image resolution and dilation status. Adjusted analyses leveraged logistic regression models fit using the statsmodels library (v0.12.1). For subgroup analysis, when a patient had multiple visits that satisfied a filtering criteria, one visit was chosen randomly. Finally, exploratory analyses were conducted for additional subgroups in validation set D (diabetes status, BMI and IOL).

Ablation analysis. We evaluated the dependence of DLS performance on the visibility of different regions of the image. Because both the images and ocular anatomy are 'circular' (that is, roughly radially symmetric), and most pupils are centred, we conducted this 'image visibility' analysis based on concentric circles. Two types of masking were examined: keeping only the centre visible and keeping the peripheral rim visible. At various degrees of masking, the two types of masking were compared while controlling for the number of visible pixels. For each condition, the same mask was applied during both training and evaluation.

Saliency analysis. From validation set A, 100 positive images with the highest predicted likelihoods were selected for saliency analysis for each task. Images were manually inspected to ensure correct vertical orientation and to orient the nasal

aspect on the right. Three saliency methods were applied: GradCAM⁶⁷, guided backprop⁶⁸ and integrated gradient⁶⁹, as implemented in the People+AI Research saliency library⁷⁰.

Image resolution analysis. When simulating the low-resolution images, images were first down-sampled to the specified resolution using the area-based method (tf.image.resize with method=AREA). Next, to ensure that the DLS input size (and, thus, the network's number of parameters and capacity) remained the same for a fair comparison, we up-sampled each image back to the original resolution (by bilinear interpolation with antialias using tf.image.resize with antialias=True). Down-sampling and up-sampling methods were chosen to produce the most visually appropriate images, blinded to the actual DLS results.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The main data supporting the results in this study are available within the paper and its Supplementary Information. This study used de-identified data from EyePACS Inc. and the teleretinal diabetes screening programme at the Atlanta Veterans Affairs. Interested researchers should contact J.C. (jcuadros@eyepacs.com) to enquire about access to EyePACS data and approach the Office of Research and Development at https://www.research.va.gov/resources/ORD_Admin/ord_contacts.cfm to enquire about access to VA data.

Code availability

The deep-learning framework (TensorFlow) used in this study is available at <https://www.tensorflow.org>; the neural network architecture is available from https://github.com/tensorflow/models/blob/master/research/slim/nets/inception_v3.py; and an ImageNet pretrained checkpoint is available from <https://github.com/tensorflow/models/tree/master/research/slim>.

Received: 30 November 2020; Accepted: 15 February 2022;

Published online: 29 March 2022

References

- Poplin, R. et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat. Biomed. Eng.* **2**, 158–164 (2018).
- Cheung, C. Y. et al. A deep-learning system for the assessment of cardiovascular disease risk via the measurement of retinal-vessel calibre. *Nat. Biomed. Eng.* <https://doi.org/10.1038/s41551-020-00626-4> (2020).
- Mitani, A. et al. Detection of anaemia from retinal fundus images via deep learning. *Nat. Biomed. Eng.* **4**, 18–27 (2020).
- Sabanayagam, C. et al. A deep-learning algorithm to detect chronic kidney disease from retinal photographs in community-based populations. *Lancet Digit. Health* **2**, e295–e302 (2020).
- Rim, T. H. et al. Prediction of systemic biomarkers from retinal photographs: development and validation of deep-learning algorithms. *Lancet Digit. Health* **2**, e526–e536 (2020).
- Tarlan, B. & Kiratli, H. Subconjunctival hemorrhage: risk factors and potential indicators. *Clin. Ophthalmol.* **7**, 1163–1170 (2013).
- Hreidarsson, A. B. Pupil size in insulin-dependent diabetes. Relationship to duration, metabolic control, and long-term manifestations. *Diabetes* **31**, 442–448 (1982).
- Smith, S. E., Smith, S. A., Brown, P. M., Fox, C. & Sonksen, P. H. Pupillary signs in diabetic autonomic neuropathy. *Br. Med. J.* **2**, 924–927 (1978).
- Banaee, T. et al. Distribution of different-sized ocular surface vessels in diabetics and normal individuals. *J. Ophthalmic Vis. Res.* **12**, 361–367 (2017).
- Iroshan, K. A. et al. Detection of diabetes by macrovascular tortuosity of superior bulbar conjunctiva. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* **2018**, 1–4 (2018).
- Comprehensive Diabetes Care. *National Committee for Quality Assurance* <https://www.ncqa.org/hedis/measures/comprehensive-diabetes-care/> (2020).
- Zangemeister, W. H., Gronow, T. & Grzyska, U. Pupillary responses to single and sinusoidal light stimuli in diabetic patients. *Neurol. Int.* **1**, e19 (2009).
- Hashemi, H. et al. White-to-white corneal diameter distribution in an adult population. *J. Curr. Ophthalmol.* **27**, 21–24 (2015).
- Rüfer, F., Schröder, A. & Erb, C. White-to-white corneal diameter: normal values in healthy humans obtained with the Orbscan II topography system. *Cornea* **24**, 259–261 (2005).
- Worthen, D. M., Fenton, B. M., Rosen, P. & Zweifach, B. Morphometry of diabetic conjunctival blood vessels. *Ophthalmology* **88**, 655–657 (1981).
- Danilova, A. I. Blood circulation in the conjunctival blood vessels of patients with diabetes mellitus. *Probl. Endokrinol.* **26**, 9–14 (1980).

17. Fenton, B. M., Zweifach, B. W. & Worthen, D. M. Quantitative morphometry of conjunctival microcirculation in diabetes mellitus. *Microvasc. Res.* **18**, 153–166 (1979).
18. Owen, C. G. et al. Diabetes and the tortuosity of vessels of the bulbar conjunctiva. *Ophthalmology* **115**, e27–e32 (2008).
19. Owen, C. G., Newsom, R. S. B., Rudnicka, A. R., Ellis, T. J. & Woodward, E. G. Vascular response of the bulbar conjunctiva to diabetes and elevated blood pressure. *Ophthalmology* **112**, 1801–1808 (2005).
20. Khan, M. A. et al. A clinical correlation of conjunctival microangiopathy with grades of retinopathy in type 2 diabetes mellitus. *Armed Forces Med. J. India* **73**, 261–266 (2017).
21. Sharma, R., Sati, A., Shankar, S. & Gurunadh, V. S. Use of conjunctival vessel width for assessment of severity of retinopathy in type-2 diabetes mellitus patients. *Int. J. Contemp. Med. Res.* **4**, 1007–1010 (2017).
22. Chang, H.-C., Sung, C.-W. & Lin, M.-H. Serum lipids and risk of atherosclerosis in xanthelasma palpebrarum: a systematic review and meta-analysis. *J. Am. Acad. Dermatol.* **82**, 596–605 (2020).
23. Piyasena, M. M. P. N. et al. Systematic review on barriers and enablers for access to diabetic retinopathy screening services in different income settings. *PLoS ONE* **14**, e0198979 (2019).
24. Stevenson, M., Lloyd-Jones, M., Morgan, M. Y. & Wong, R. in *Non-Invasive Diagnostic Assessment Tools for the Detection of Liver Fibrosis in Patients with Suspected Alcohol-Related Liver Disease: A Systematic Review and Economic Evaluation* Appendix 8 (NIHR Journals Library, 2012).
25. Bang, H. et al. Development and validation of a patient self-assessment score for diabetes risk. *Ann. Intern. Med.* **151**, 775–783 (2009).
26. American Diabetes Association and the Centers for Disease Control and Prevention. Prediabetes Risk Test. *National Diabetes Prevention Program* <https://www.cdc.gov/diabetes/prevention/pdf/Prediabetes-Risk-Test-Final.pdf> (2009).
27. Zhang, X. et al. Access to health care and control of ABCs of diabetes. *Diabetes Care* **35**, 1566–1571 (2012).
28. Klonoff, D. C. & Schwartz, D. M. An economic analysis of interventions for diabetes. *Diabetes Care* **23**, 390–404 (2000).
29. Keenum, Z. et al. Patients' adherence to recommended follow-up eye care after diabetic retinopathy screening in a publicly funded county clinic and factors associated with follow-up eye care use. *JAMA Ophthalmol.* **134**, 1221–1228 (2016).
30. Lu, Y. et al. Divergent perceptions of barriers to diabetic retinopathy screening among patients and care providers, Los Angeles, California, 2014–2015. *Prev. Chronic Dis.* **13**, E140 (2016).
31. Paz, S. H. et al. Noncompliance with vision care guidelines in Latinos with type 2 diabetes mellitus: the Los Angeles Latino Eye Study. *Ophthalmology* **113**, 1372–1377 (2006).
32. Legorreta, A. P., Hasan, M. M., Peters, A. L., Pelletier, K. R. & Leung, K. M. An intervention for enhancing compliance with screening recommendations for diabetic retinopathy. A bicoastal experience. *Diabetes Care* **20**, 520–523 (1997).
33. Nelson, R. H. Hyperlipidemia as a risk factor for cardiovascular disease. *Prim. Care* **40**, 195–211 (2013).
34. Karr, S. Epidemiology and management of hyperlipidemia. *Am. J. Manag. Care* **23**, S139–S148 (2017).
35. Alexander, G. C. et al. Use and content of primary care office-based vs telemedicine care visits during the COVID-19 pandemic in the US. *JAMA Netw. Open* **3**, e2021476 (2020).
36. Jalil, M., Ferenczy, S. R. & Shields, C. L. iPhone 4s and iPhone 5s imaging of the eye. *Ocul. Oncol. Pathol.* **3**, 49–55 (2017).
37. Ludwig, C. A. et al. Training time and quality of smartphone-based anterior segment screening in rural India. *Clin. Ophthalmol.* **11**, 1301–1307 (2017).
38. Avram, R. et al. A digital biomarker of diabetes from smartphone-based vascular signals. *Nat. Med.* **26**, 1576–1582 (2020).
39. Santos, M. & Hofmann, R. J. Ocular manifestations of obstructive sleep apnea. *J. Clin. Sleep Med.* **13**, 1345–1348 (2017).
40. Cristescu Teodor, R. & Mihaltan, F. D. Eyelid laxity and sleep apnea syndrome: a review. *Rom. J. Ophthalmol.* **63**, 2–9 (2019).
41. Scott, I. U. & Siatkowski, M. R. Thyroid eye disease. *Semin. Ophthalmol.* **14**, 52–61 (1999).
42. Dutton, J. J. Anatomic considerations in thyroid eye disease. *Ophthalm. Plast. Reconstr. Surg.* **34**, S7–S12 (2018).
43. Christoffersen, M. et al. Xanthelasmata, arcus corneae, and ischaemic vascular disease and death in general population: prospective cohort study. *Br. Med. J.* **343**, d5497 (2011).
44. To, W. J. et al. Real-time studies of hypertension using non-mydratric fundus photography and computer-assisted intravital microscopy. *Clin. Hemorheol. Microcirc.* **53**, 267–279 (2013).
45. Mullaem, G. & Rosner, M. H. Ocular problems in the patient with end-stage renal disease. *Semin. Dial.* **25**, 403–407 (2012).
46. Klaassen-Broekema, N. & van Bijsterveld, O. P. Limbal and corneal calcification in patients with chronic renal failure. *Br. J. Ophthalmol.* **77**, 569–571 (1993).
47. Sharon, Y. & Schlesinger, N. Beyond joints: a review of ocular abnormalities in gout and hyperuricemia. *Curr. Rheumatol. Rep.* **18**, 37 (2016).
48. Hsiao, C.-H. et al. Association of severity of conjunctival and corneal calcification with all-cause 1-year mortality in maintenance haemodialysis patients. *Nephrol. Dial. Transpl.* **26**, 1016–1023 (2011).
49. Xiao, W. et al. Screening and identifying hepatobiliary diseases through deep learning using ocular images: a prospective, multicentre study. *Lancet Digit. Health* **3**, e88–e97 (2021).
50. Wagner, S. K. et al. Insights into systemic disease through retinal imaging-based ophthalmics. *Transl. Vis. Sci. Technol.* **9**, 6 (2020).
51. Panwar, N. et al. Fundus photography in the 21st century — a review of recent technological advances and their implications for worldwide healthcare. *Telemed. J. E. Health* **22**, 198–208 (2016).
52. Prokofyeva, E., Wegener, A. & Zrenner, E. Cataract prevalence and prevention in Europe: a literature review. *Acta Ophthalmol.* **91**, 395–405 (2013).
53. Madsen, P. H. Rubeosis of the iris and haemorrhagic glaucoma in patients with proliferative diabetic retinopathy. *Br. J. Ophthalmol.* **55**, 368–371 (1971).
54. Rodrigues, G. B. et al. Neovascular glaucoma: a review. *Int. J. Retin. Vitreous* **2**, 26 (2016).
55. Zhou, Y., Zhang, Y., Shi, K. & Wang, C. Body mass index and risk of diabetic retinopathy: a meta-analysis and systematic review. *Medicine* **96**, e6754 (2017).
56. Babic, B., Gerke, S., Evgeniou, T. & Glenn Cohen, I. Direct-to-consumer medical machine learning and artificial intelligence applications. *Nat. Mach. Intell.* **3**, 283–287 (2021).
57. Cuadros, J. & Bresnick, G. EyePACS: an adaptable telemedicine system for diabetic retinopathy screening. *J. Diabetes Sci. Technol.* **3**, 509–516 (2009).
58. Chasan, J. E., Delaune, B., Maa, A. Y. & Lynch, M. G. Effect of a teleretinal screening program on eye care use and resources. *JAMA Ophthalmol.* **132**, 1045–1051 (2014).
59. Maa, A. Y. et al. Early experience with technology-based eye care services (TECS): a novel ophthalmologic telemedicine initiative. *Ophthalmology* **124**, 539–546 (2017).
60. *Photographer Manual* (EyePACS); https://www.eyepacs.org/photographer/protocol.jsp#external_photos (2008).
61. 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* **98**, 786–806 (1991).
62. *EyePACS digital retinal grading protocol* (EyePACS); <https://www.eyepacs.org/consultant/Clinical/grading/EyePACS-DIGITAL-RETINAL-IMAGE-GRADING.pdf> (2008).
- AQ4163. Bora, A. et al. Predicting risk of developing diabetic retinopathy using deep learning. *Lancet Digit. Health* **3**, e10–e19 (2021).
64. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2818–2826 (IEEE, 2016).
65. 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).
66. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (1988).
67. Selvaraju, R. R. et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)* 618–626 (IEEE, 2017).
68. Springenberg, J. T., Dosovitskiy, A., Brox, T. & Riedmiller, M. Striving for Simplicity: The All Convolutional Net. Preprint at *arXiv [cs.LG]* <https://arxiv.org/abs/1412.6806> (2014).
69. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning* Vol. 70, 3319–3328 (2017).
70. PAIR-code. <https://github.com/PAIR-code/saliency> (2020).

Acknowledgements

This work was funded by Google LLC. We acknowledge H. Doan, Q. Duong, R. Lee and the Google Health team for software infrastructure support and data collection. We also thank T. Guo, M. McConnell, M. Howell and S. Kavusi for their feedback on the manuscript. We are grateful to the graders who labelled data for the pupil segmentation model.

Author contributions

B.B., A.M. and N.K. conducted the machine-learning model development, experiments and statistical analysis with input and guidance from A.V., N.H. and Y.L. B.B., A.M., I.T., N.H. and Y.L. designed the study and pre-specified the statistical analysis. I.T., N.K. and P.S. developed guidelines and managed data collection for the pupil/iris segmentation model. J.C. and A.Y.M. managed data collection and associated approvals. G.S.C., L.P. and D.R.W. obtained funding for data collection and analysis, supervised the study and provided strategic guidance. B.B., A.M., I.T., N.H. and Y.L. prepared the manuscript with input from all authors. B.B. and A.M. contributed equally, and A.V., N.H. and Y.L. contributed equally.

Competing interests

B.B., A.M., N.K., G.S.C., L.P., D.R.W., A.V., N.H. and Y.L. are employees of Google LLC, own Alphabet stock and are co-inventors on patents (in various stages) for machine learning using medical images. I.T. is a consultant of Google LLC. J.C. is the CEO of EyePACS Inc. A.Y.M. declares no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41551-022-00867-5>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41551-022-00867-5>.

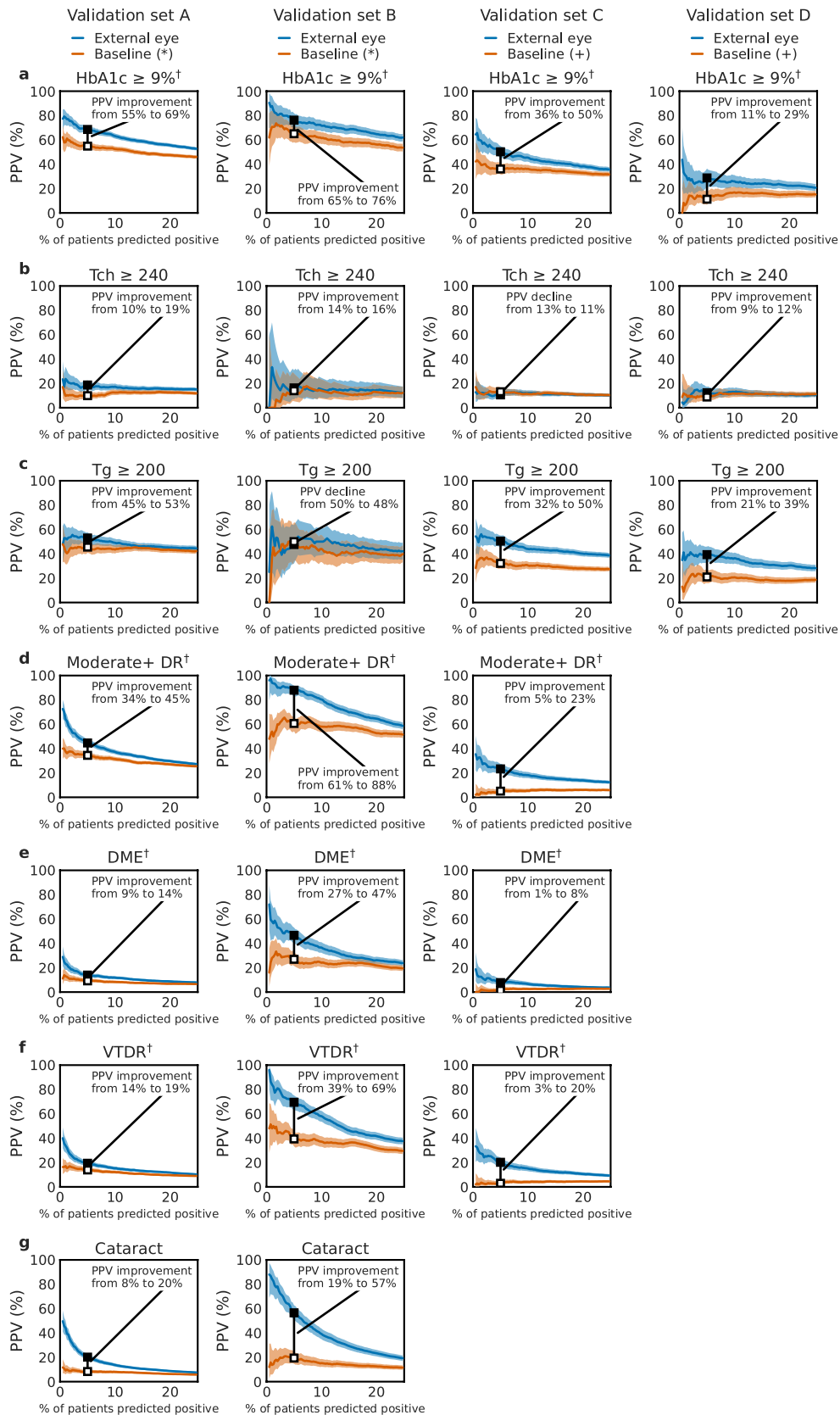
Correspondence and requests for materials should be addressed to Naama Hammel or Yun Liu.

Peer review information *Nature Biomedical Engineering* thanks Meindert Niemeijer and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

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

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

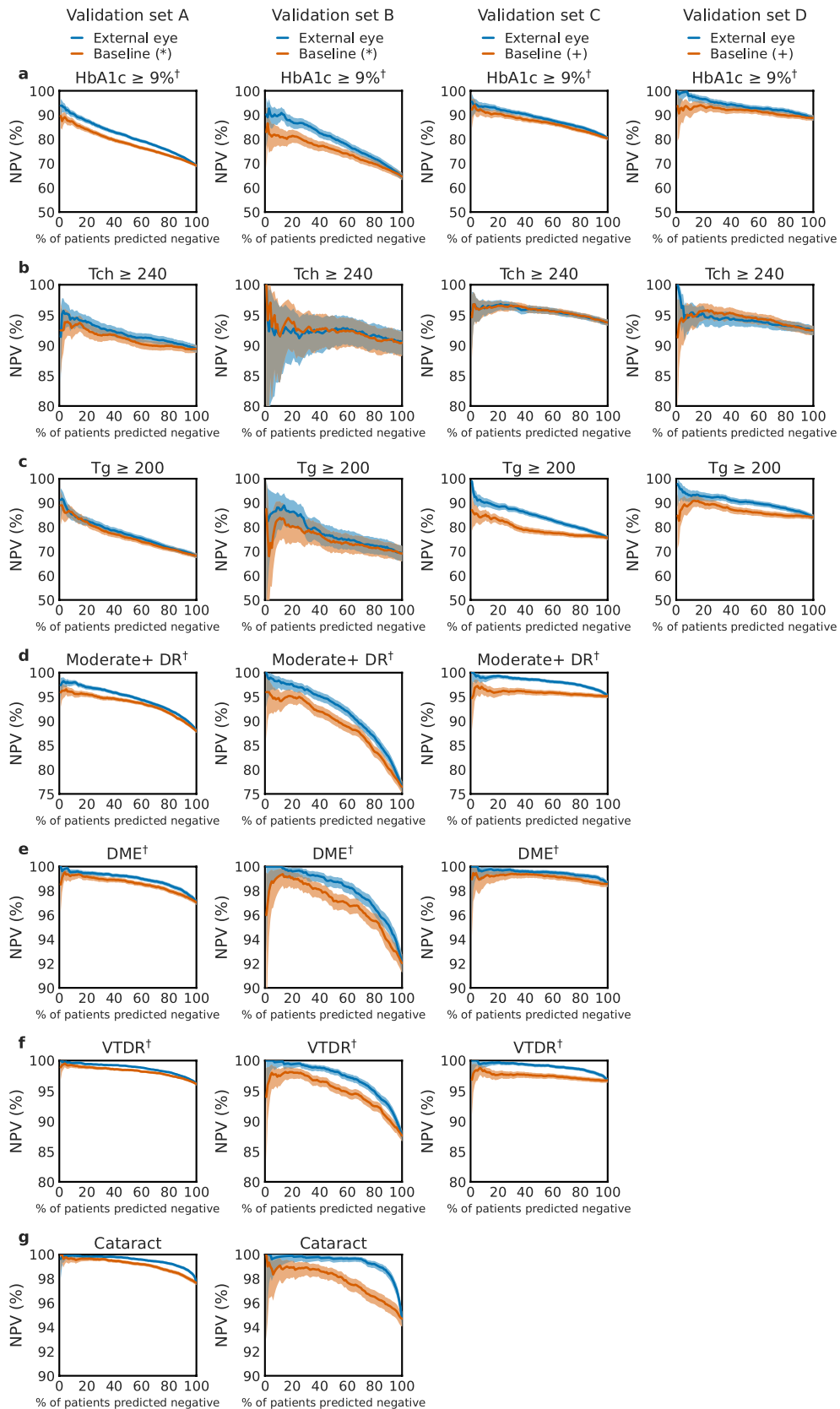
© The Author(s), under exclusive licence to Springer Nature Limited 2022



Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Curves of positive predictive value (PPV) as a function of threshold for various predictions using external eye images.

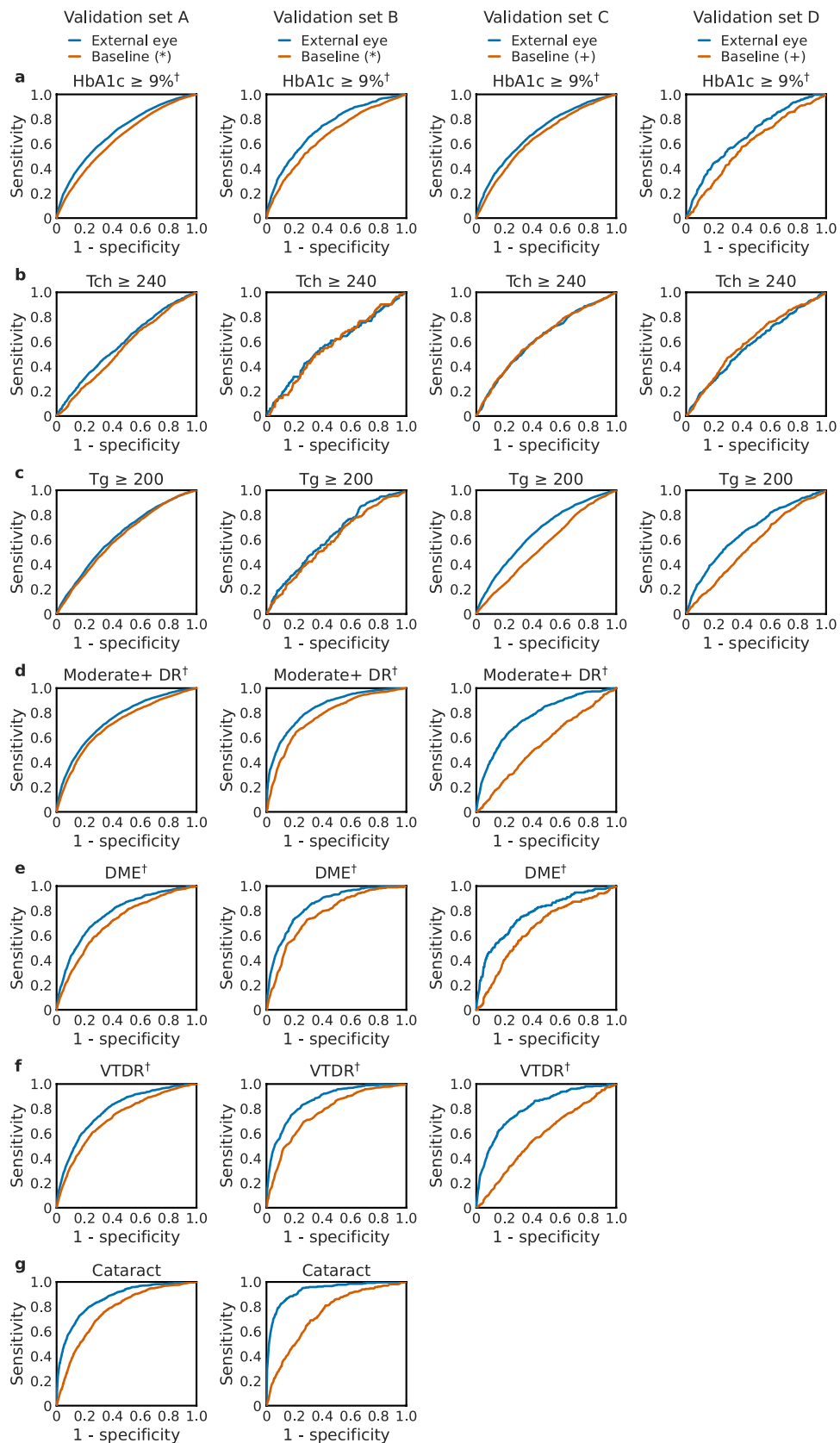
a, poor sugar control ($\text{HbA1c} \geq 9\%$), **b-c**, elevated lipids (total cholesterol $\geq 240 \text{ mg dl}^{-1}$ and triglycerides $\geq 200 \text{ mg dl}^{-1}$), **d**, moderate-or-worse diabetic retinopathy (DR), **e**, diabetic macular edema (DME), **f**, vision-threatening DR (VTDR), and **g**, a positive control: cataract. In these plots, the x-axis indicates the percentage of patients predicted to be positive; for example 5% means the top 5% based on predicted likelihood was categorized to be “positive”, and the respective curves indicate the PPV for that threshold. The curves are truncated at the extreme end (when only 0.5% of patients are predicted positive, confidence intervals are wide) to reduce noise and improve clarity. Shaded areas indicate 95% bootstrap confidence intervals. Empty panels indicate unavailable data in validation set C and D. (*) Baseline characteristics models for validation sets A and B include self-reported age, sex, race/ethnicity and years with diabetes and were trained on the training dataset. (+) The baseline characteristics models for validation sets C and D use self-reported age and sex and were trained directly on validation sets C and D due to large differences in patient population compared to the development set. †: prespecified primary prediction tasks.



Extended Data Fig. 2 | See next page for caption.

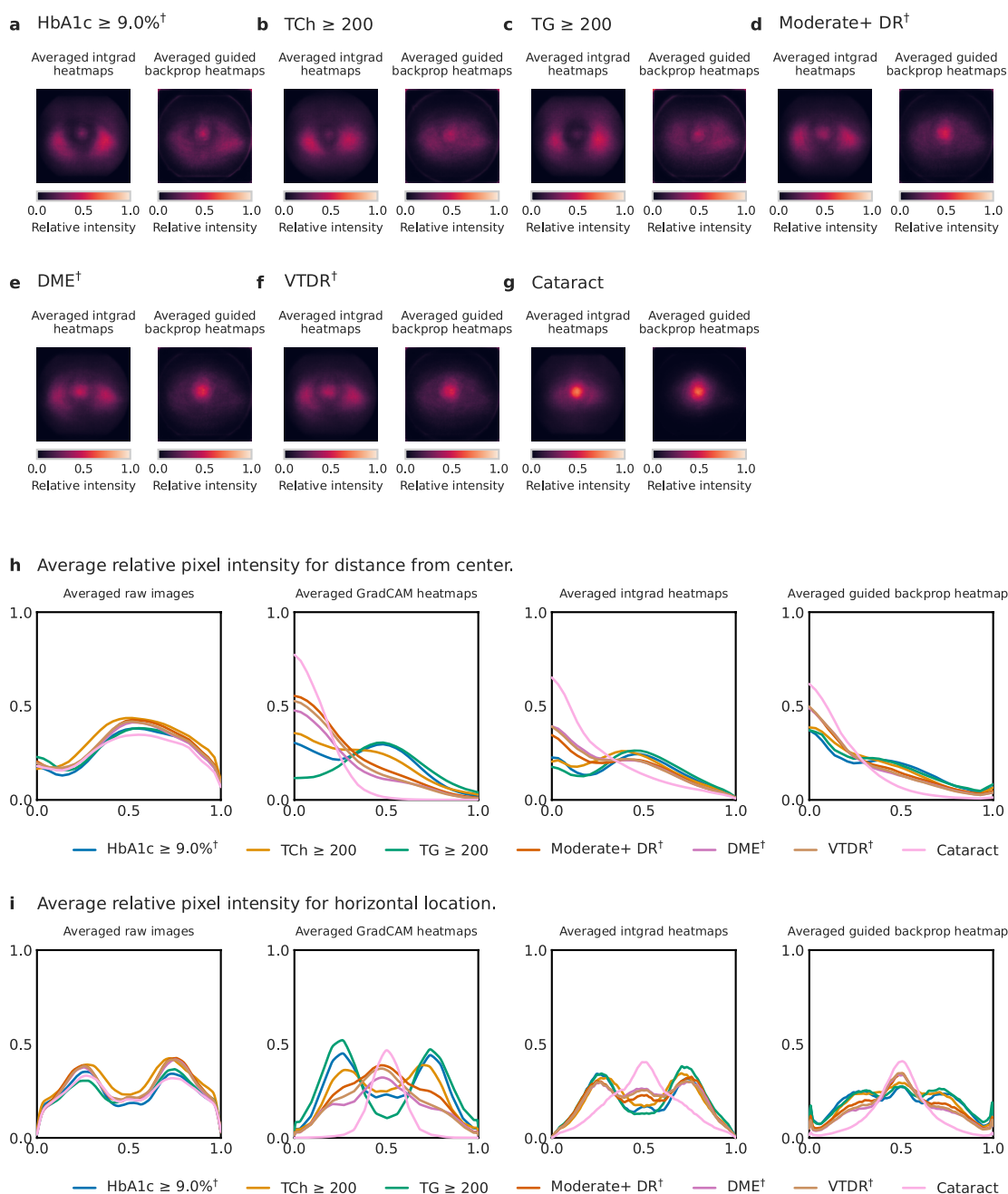
Extended Data Fig. 2 | Curves of negative predictive values (NPV) as a function of threshold for various predictions using external eye images.

a, poor sugar control ($\text{HbA1c} \geq 9\%$), **b-c**, elevated lipids (total cholesterol $\geq 240 \text{ mg dl}^{-1}$ and triglycerides $\geq 200 \text{ mg dl}^{-1}$), **d**, moderate-or-worse diabetic retinopathy (DR), **e**, diabetic macular edema (DME), **f**, vision-threatening DR (VTDR), and **g**, a positive control: cataract. This is the NPV equivalent of Extended Data Fig. 1. The curves are truncated at the extreme end (when only 1% of patients are predicted negative, confidence intervals are wide) to reduce noise and improve clarity. Shaded areas indicate 95% bootstrap confidence intervals. Empty panels indicate unavailable data in validation set C and D. (*) Baseline characteristics models for validation sets A and B include self-reported age, sex, race/ethnicity and years with diabetes and were trained on the training dataset. (+) The baseline characteristics models for validation sets C and D use self-reported age and sex and were trained directly on validation sets C and D due to large differences in patient population compared to the development set. †: prespecified primary prediction tasks.



Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Receiver operating characteristic curves (ROCs) for various predictions using external eye images. **a**, poor sugar control (HbA1c $\geq 9\%$), **b-c**, elevated lipids (total cholesterol ≥ 240 mg dl⁻¹ and triglycerides ≥ 200 mg dl⁻¹), **d**, moderate-or-worse diabetic retinopathy (DR), **e**, diabetic macular edema (DME), **f**, vision-threatening DR (VTDR), and **g**, a positive control: cataract. Sample sizes ("N" for the number of visits and "n" for the number of positive visits), area under ROC (AUCs), and the p-value for the difference are provided in Supplementary Table 1. Empty panels indicate unavailable data in validation sets C and D. (*) Baseline characteristics models for validation sets A and B include self-reported age, sex, race/ethnicity and years with diabetes and were trained on the training dataset. (+) The baseline characteristics models for validation sets C and D use self-reported age and sex and were trained directly on validation sets C and D due to large differences in patient population compared to the development set. †: prespecified primary prediction tasks.



Extended Data Fig. 4 | Saliency analysis illustrating the influence of various regions of the image towards the prediction. a-g, Figures are generated in the same manner as in Fig. 3, but with different saliency methods: Integrated Gradients on the left, and guided backpropagation on the right. **h-i,** Quantifying the pixel intensity in the averaged saliency heatmaps. † : prespecified primary prediction tasks.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- | | | |
|-------------------------------------|-------------------------------------|--|
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | A description of all covariates tested |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection No software was used.

Data analysis The open-sourced library TensorFlow was used to develop the models, and the statsmodels library and custom Python code were used for statistical analysis.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

The main data supporting the results in this study are available within the paper and its Supplementary Information. This study used de-identified data from EyePACS Inc. and the teleretinal diabetes screening program at the Atlanta Veterans Affairs. Interested researchers should contact J.C. (jcuadros@eyepacs.com) to inquire about access to EyePACS data and approach the Office of Research and Development at https://www.research.va.gov/resources/ORD_Admin/ord_contacts.cfm to inquire about access to VA data.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Explicit sample-size calculations were not done; instead, we aimed at having geographically separate regions for broad external validation (that is, if the model was trained on California data, it was evaluated on non-California data).
Data exclusions	The images and baseline characteristics needed to be available. In addition, Los Angeles county was excluded from the training data because it was held out for other projects.
Replication	Broad geographical validation across 4 validation sets spanning 198 sites in 18 other US states.
Randomization	Randomization was not applicable, because this study was not interventional. However, a negative control prediction (cataract) was used for the saliency experiments.
Blinding	All grading or labelling were done blinded to model predictions.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involvement
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input type="checkbox"/>	<input checked="" type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Involvement
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	Details are provided in Table 1. Briefly, the EyePACS datasets and validation set C correspond to patients with diabetes, whereas validation set D also includes patients without diabetes who presented to eye care at the VA.
Recruitment	This was a retrospective study involving de-identified data, and hence there was no explicit enrollment. Please see Methods for details.
Ethics oversight	Ethics review and Institutional Review Board exemptions for this retrospective study on de-identified data were obtained via the Advarra Review Institutional Review Board.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about [clinical studies](#)

All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

Clinical trial registration	Not applicable, because this study was not prospective nor interventional.
Study protocol	Not applicable, because this study was not prospective nor interventional.

Data collection

This work used de-identified images and baseline characteristics from 4 teleretinal eye screening datasets in the US. The first two datasets were the California and non-California cohorts from EyePACS, a teleretinal diabetic screening service in the US. The third and fourth datasets were from the Atlanta Veterans Affairs (VA), which served multiple community-based outpatient clinics (CBOCs) in the greater Atlanta area.

Outcomes

HgA1c values were extracted from the relevant records by each site; diabetic retinal diseases were graded by graders blinded to model predictions.