JCI The Journal of Clinical Investigation

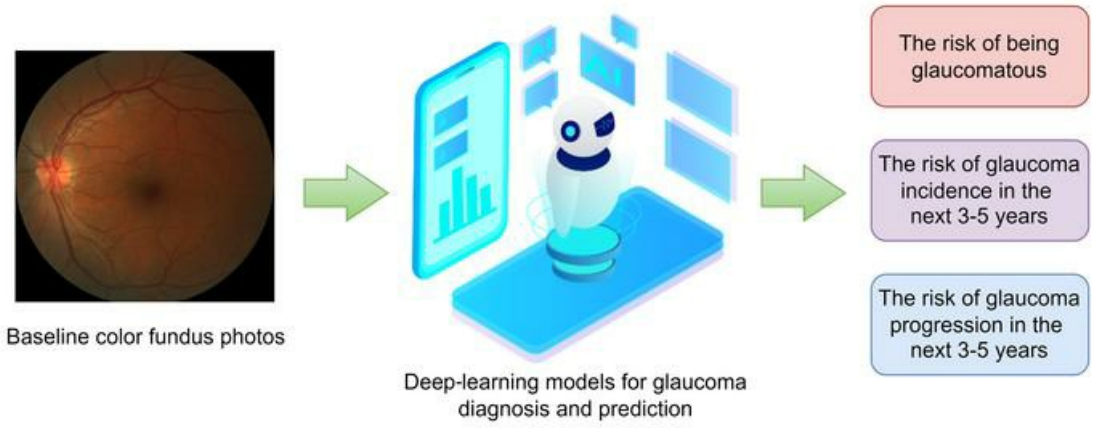
A deep-learning system predicts glaucoma incidence and progression using retinal photographs

Fei Li, … , Kang Zhang, Xiulan Zhang

J Clin Invest. 2022;132(11):e157968. https://doi.org/10.1172/JCI157968.

Clinical Medicine | Ophthalmology

Graphical abstract



A deep-learning system predicts glaucoma incidence and progression using retinal photographs

Fei Li,[1] Yuandong Su,[2,3] Fengbin Lin,[1] Zhihuan Li,[3] Yunhe Song,[1] Sheng Nie,[4] Jie Xu,[5] Linjiang Chen,[6] Shiyan Chen,[7] Hao Li,[8] Kanmin Xue,[9] Huixin Che,[10] Zhengui Chen,[11] Bin Yang,[12] Huiying Zhang,[13] Ming Ge,[14] Weihui Zhong,[15] Chunman Yang,[16] Lina Chen,[17] Fanyin Wang,[18] Yunqin Jia,[19] Wanlin Li,[20] Yuqing Wu,[21] Yingjie Li,[22] Yuanxu Gao,[3,23] Yong Zhou,[24] Kang Zhang,[3] and Xiulan Zhang[1]

[1]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science, Guangdong Provincial Clinical Research Center for Ocular Diseases, Guangzhou, China. [2]State Key Laboratory of Biotherapy and Center for Translational Innovations, West China Hospital and Sichuan University, Chengdu, China. [3]PKU-MUST Center for Future Technology, Faculty of Medicine, Macao University of Science and Technology, Macau, China. [4]State Key Laboratory of Organ Failure Research, National Clinical Research Center for Kidney Disease and Nanfang Hospital, Southern Medical University, Guangzhou, China. [5]Beijing Institute of Ophthalmology, Beijing Tongren Eye Center, Beijing Tongren Hospital, Beijing Ophthalmology and Visual Science Key Lab, Beijing, China. [6]Department of Ophthalmology, Nanfang Hospital, Southern Medical University, Guangzhou, China. [7]Department of Ophthalmology, Sichuan Academy of Medical Sciences & Sichuan Provincial People's Hospital, Chengdu, China. [8]Department of Ophthalmology, Guizhou Provincial People's Hospital, Guiyang, China. [9]Nuffield Laboratory of Ophthalmology, Department of Clinical Neurosciences, University of Oxford and Oxford University Hospitals NHS Foundation Trust, Oxford, United Kingdom. [10]He Eye Specialist Hospital, Shenyang, Liaoning Province, China. [11]Jiangmen Xinhui Aier New Hope Eye Hospital, Jiangmen, Guangdong, China. [12]Department of Ophthalmology, Zigong Third People's Hospital, Zigong, China. [13]Department of Ophthalmology, Fujian Provincial Hospital, Fuzhou, China. [14]Department of Ophthalmology and Optometry, Guizhou Nursing Vocational College, Guiyang, China. [15]Department of Ophthalmology, Guangzhou Development District Hospital, Guangzhou, China. [16]Department of Ophthalmology, The Second Affiliated Hospital of Guizhou Medical University, Kaili, China. [17]Department of Ophthalmology, The Third People's Hospital of Dalian, Dalian, Liaoning Province, China. [18]Department of Ophthalmology, Shenzhen Qianhai Shekou Free Trade Zone Hospital, Shenzhen, China. [19]Department of Ophthalmology, Dali Bai Autonomous Prefecture People's Hospital, Dali, China. [20]Department of Ophthalmology, Wuwei People's Hospital, Wuwei, Gansu Province, China. [21]Department of Ophthalmology, Joint Shantou International Eye Center of Shantou University and the Chinese University of Hong Kong, Shantou, Guangdong, China. [22]Department of Ophthalmology, The First Hospital of Nanchang City, Nanchang, China. [23]State Key Laboratory of Lunar and Planetary Sciences, Macao University of Science and Technology, Taipa, Macau, China. [24]Clinical Research Institute, Shanghai General Hospital, Shanghai Jiaotong University School of Medicine, Shanghai, China.

BACKGROUND. Deep learning has been widely used for glaucoma diagnosis. However, there is no clinically validated algorithm for glaucoma incidence and progression prediction. This study aims to develop a clinically feasible deep-learning system for predicting and stratifying the risk of glaucoma onset and progression based on color fundus photographs (CFPs), with clinical validation of performance in external population cohorts.

METHODS. We established data sets of CFPs and visual fields collected from longitudinal cohorts. The mean follow-up duration was 3 to 5 years across the data sets. Artificial intelligence (AI) models were developed to predict future glaucoma incidence and progression based on the CFPs of 17,497 eyes in 9346 patients. The area under the receiver operating characteristic (AUROC) curve, sensitivity, and specificity of the AI models were calculated with reference to the labels provided by experienced ophthalmologists. Incidence and progression of glaucoma were determined based on longitudinal CFP images or visual fields, respectively.

RESULTS. The AI model to predict glaucoma incidence achieved an AUROC of 0.90 (0.81–0.99) in the validation set and demonstrated good generalizability, with AUROCs of 0.89 (0.83–0.95) and 0.88 (0.79–0.97) in external test sets 1 and 2, respectively. The AI model to predict glaucoma progression achieved an AUROC of 0.91 (0.88–0.94) in the validation set, and also demonstrated outstanding predictive performance with AUROCs of 0.87 (0.81–0.92) and 0.88 (0.83–0.94) in external test sets 1 and 2, respectively.

CONCLUSION. Our study demonstrates the feasibility of deep-learning algorithms in the early detection and prediction of glaucoma progression.

FUNDING. National Natural Science Foundation of China (NSFC); the High-level Hospital Construction Project, Zhongshan Ophthalmic Center, Sun Yat-sen University; the Science and Technology Program of Guangzhou, China (2021), the Science and Technology Development Fund (FDCT) of Macau, and FDCT-NSFC.

Introduction

Glaucoma is a major chronic eye disease characterized by optic nerve damage and visual field defects (1, 2). Its onset is often insidious, with a risk of irreversible visual field loss prior to becoming symptomatic (3). Timely detection and treatment of glaucoma by lowering the intraocular pressure (IOP) could reduce the risk of disease progression (4, 5). Predicting glaucoma onset and progression

Authorship note: F Li, Y Su, F Lin, and ZL contributed equally to this work.
Conflict of interest: The authors have declared that no conflict of interest exists.

Submitted: December 29, 2021; **Accepted:** April 12, 2022; **Published:** June 1, 2022.
Reference information: *J Clin Invest*. 2022;132(11):e157968.
https://doi.org/10.1172/JCI157968.

is a major clinical challenge. Previous studies demonstrated that biometric parameters, such as baseline IOP, vertical cup-to-disc ratio, mean deviation (in the Humphrey visual field test), and pattern standard deviation are helpful in predicting glaucoma incidence and progression (6–12). However, IOP measurement and visual field tests are often not available in the primary healthcare setting. In contrast, color fundus photography is widely available and color fundus photographs (CFPs) can be rapidly acquired, with the potential to allow artificial intelligence–based (AI-based) diagnosis of optic nerve, retinal, and systemic diseases (including chronic kidney disease, diabetes mellitus; ref. 13). Smartphones can also be adapted to capture CFPs, making them a promising tool in disease screening in the future (14, 15). Thus, it would be advantageous if glaucoma incidence and progression could be solely based on CFPs rather than relying on multiple test modalities.

Deep-learning techniques have been widely used for glaucoma diagnosis (16–19). However, there is no clinically validated algorithm for glaucoma incidence and progression prediction. This study aimed to develop a clinically feasible deep-learning system for diagnosing glaucoma (Figure 1, A and B) and predicting the risk of glaucoma onset and progression (Figure 1, C and D) based on CFPs, with validation of performance in external population cohorts. Our AI system appears to be capable of detecting features in the baseline CFPs that are unrecognizable to the human eye and predict which patients will progress to glaucoma within 5 years. Furthermore, we show that the AI system could be deployed at the point of care via smartphone image capture to enable broadly accessible remote glaucoma screening in the future.

Results

Definitions of glaucoma, its incidence, and progression. The diagnostic criteria for possible glaucoma based on CFPs were created following published population-based studies; glaucomatous optic neuropathy was defined by the presence of a vertical cup-to-disc ratio of 0.7 or greater, retinal nerve fiber layer (RNFL) defect, optic disc rim width of 0.1-disc diameter or smaller, and/or disc hemorrhage (20–22). Glaucoma incidence was defined as eyes having nonglaucomatous baseline CFPs but becoming possibly glaucomatous during a follow-up period.

Humphrey visual fields performed in a standard 24-2 pattern mode were used for an analysis when glaucoma progression was suspected (23). Glaucomatous progression was defined by at least 3 visual field test points worse than the baseline at the 5% level in 2 consecutive reliable visual field tests or at least 3 visual field locations worse than the baseline at the 5% level in 2 subsequent consecutive reliable visual field tests (23). Time to progression was defined as the time from a baseline to the first visual field test report that confirmed glaucoma progression following the aforementioned criteria. The gold standard definition of clinical progression was confirmed to have been met by unanimous agreement of 3 ophthalmologists who independently assessed each visual field report.

Image data sets and patient characteristics. We established a large data set composed of CFPs and visual fields collected in Guangzhou, Beijing, and Kashi, China. The demographic and clinical information of the study participants is summarized in Table 1. The data were split randomly into mutually exclusive sets for training, validation, and external testing of the AI algorithms.

In the first task, we developed a model to diagnose possible glaucoma based on 31,040 CFPs. In this task, 31,040 images (split into 20,872 for training, 3182 for validation, 6162 for external test 1,and 824 for external test 2) from 14,905 individuals were collected from glaucoma and anterior segment disease eye clinics. Among these images, 10,175 (32.8%) were diagnosed with possible glaucoma. The training and validation data sets were obtained from individuals from glaucoma and anterior segment disease sections in the Zhongshan Ophthalmic Center in Guangzhou, China. External test set 1 was collected from patients in the glaucoma and anterior segment disease clinic in Jidong Hospital near Beijing. To further test the generalizability of the AI model, we validated its performance with CFPs obtained by smartphones from Kashi.

In the second task, we developed a model to predict future glaucoma incidence based on the data from 3 longitudinal cohorts. We included a total of 13,222 eyes (10,357 training, 1191 validation, 955 external test 1, 719 external test 2) of 7127 participants, all of which were diagnosed as nonglaucomatous at the baseline. The training and validation data sets were obtained from individuals who underwent an annual health check in Guangzhou, while external test set 1 was from individuals who underwent an annual health check in Beijing and external test set 2 was from a community cohort in Guangzhou. The mean follow-up duration was 47.8–56.6 months across the data sets. The incidence rate of glaucoma was 1.1%–2.0% across the data sets.

In the third task, we developed a model to predict glaucoma progression based on the CFPs from cohorts with existing glaucoma. In this task, 4275 eyes (3003 training, 422 validation, 337 external test 1, 513 external test 2) from 2219 glaucoma patients were included, all of which were already diagnosed with glaucomatous optic neuropathy at the baseline. The training and validation data sets were obtained from 1 primary open-angle glaucoma (POAG) cohort in the Zhongshan Ophthalmic Center. To further test the generalizability of the AI model on different subtypes of glaucoma, external test set 1 was collected from another POAG cohort and external test set 2 was collected from a chronic primary angle-closure glaucoma (PACG) cohort in the Zhongshan Ophthalmic Center. The mean follow-up duration was 34.8–41.7 months across the data sets, and the proportion of glaucoma progression was 6%–13.5% across the data sets (Table 1).

Design of the diagnostic (DiagnoseNet) and predictive (PredictNet) algorithms. First, we developed a diagnostic algorithm for possible glaucoma, DiagnoseNet (Figure 1B). In brief, DiagnosetNet is composed of 2 main modules, a segmentation module and a diagnostic module. The CFPs were semantically segmented by the segmentation module with 4 anatomical structures: retinal vessels, macula, optic cup, and optic disk. The diagnostic module generated the glaucomatous probability score.

We then designed a pipeline, PredictNet, for incidence and progression prediction of glaucoma. In brief, PredictNet is also composed of 2 main modules, the segmentation module and the prediction module. The segmentation module is the same as that in DiagnoseNet. The prediction module produces the risk score of glaucoma incidence or progression in the future (Figure 1D and Supplemental Figure 1).

The diagnostic and predictive algorithms share the same segmentation module. The segmentation module was trained based

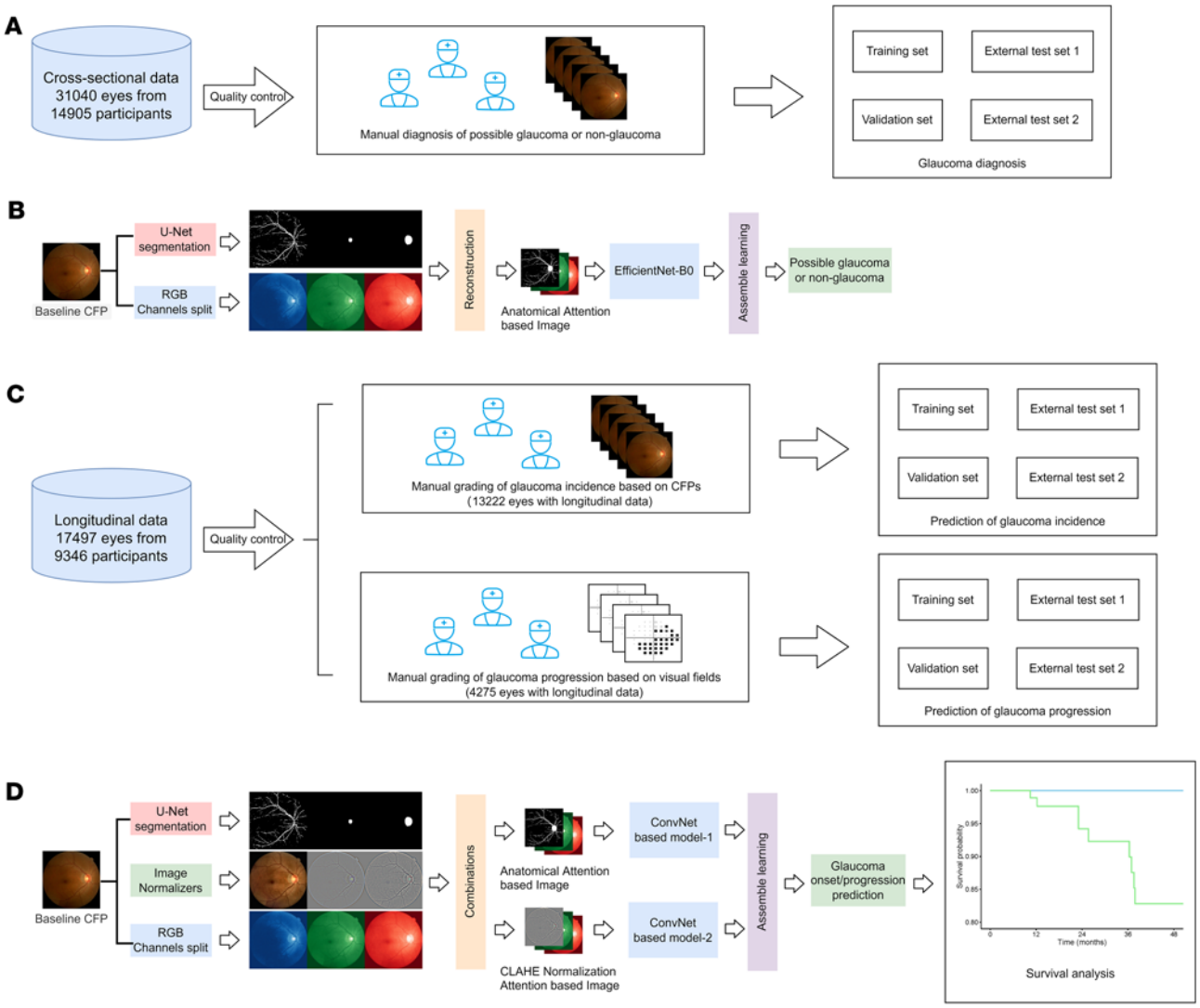


Figure 1. Development and validation of the deep-learning system for glaucoma diagnosis and incidence and progression prediction. (**A**) Data collection and ground truth labeling of glaucoma diagnosis based on CFPs. (**B**) Pipeline for glaucoma diagnosis. (**C**) Data collection and ground truth labeling of glaucoma incidence and progression. (**D**) Pipeline for predicting glaucoma development and progression. CFP, color fundus photograph; VF, visual field.

on manual annotations of optic disc (1853 images), optic cup (1860 images), macula (1695 images), and blood vessels (160 images) independently. The segmentation module demonstrated outstanding segmentation performance on the above anatomical structures and achieved an intersection over union (IOU) of 0.847, 0.669, 0.570, and 0.538 for optic disc, optic cup, macula, and blood vessel segmentation, respectively (Supplemental Table 1). Representative samples of segmentation are shown in Supplemental Figure 2.

Diagnostic performance of the AI model based on CFPs captured by smartphones. To demonstrate the potential of deploying our AI model in routine healthcare, we developed and tested the AI model to diagnose possible glaucoma based on CFPs not only from fundus cameras but also from smartphones. As shown in Table 2, in this validation data set, the AI model achieved an area under the receiver operating characteristic (AUROC) curve of 0.97 (0.96–0.97), a sensitivity of 0.98 (0.97–0.99), and a specificity of 0.82 (0.80–0.83) for differentiating glaucomatous and nonglaucomatous eyes. To evaluate the generalizability of the algorithms, the AI model was tested on 2 external data sets. In external test set 1, the AI model achieved an AUROC of 0.94 (0.93–0.94), a sensitivity of 0.89 (0.87–0.90), and a specificity of 0.83 (0.81–0.84). In external test set 2, which was obtained using smartphones, the AI model achieved an AUROC of 0.91 (0.89–0.93), a sensitivity of 0.92 (0.88–0.96), and a specificity of 0.71 (0.67–0.74).

Prediction of glaucoma incidence using longitudinal cohorts. We investigated the predictive performance of the AI model for the development of glaucoma in nonglaucomatous individuals over a 4- to 5-year period. A total of 158 eyes developed glaucoma within the 4- to 5-year period. The AI model achieved an AUROC of 0.90 (0.81–0.99), a sensitivity of 0.84 (0.82–0.87), and a specificity of 0.82 (0.57–0.96) for predicting glaucoma incidence in the validation set (Table 2 and Figure 2). The AI model demonstrated good generalizability in the external test sets, which achieved an AUROC of 0.89 (0.83–0.95), a sensitivity of 0.84 (0.81–0.86), and a specificity of 0.68 (0.43–0.87) in external test set 1, and an AUROC of 0.88 (0.79–0.97), a sensitivity of 0.84 (0.81–0.86),

Table 1. Baseline characteristics of the study participants in the different data sets

Glaucoma diagnosis	Training set	Validation set	External test set 1	External test set 2	P_1^A	P_2^B
Number of participants	10,466	1418	2866	155	-	-
Number of eyes	20,872	3182	6162	824	-	-
Male/Female	2570/7896	351/1067	732/2134	68/87	0.28	<0.0001
Age, years	54.0 (17.5)	54.9 (16.7)	61.1 (21.1)	68.8 (15.8)	0.36	<0.0001
Hypertension, %	604 (5.8%)	52 (3.7%)	135 (4.7%)	101 (64.7%)	0.03	<0.0001
Diabetes, %	989 (9.4%)	84 (5.9%)	197 (6.9%)	141 (90.4%)	<0.001	<0.0001
Number of glaucomatous eyes, %	7034 (33.7%)	952 (29.9%)	2037 (33.1%)	152 (18.4%)	0.35	<0.0001
Glaucoma onset prediction	**Training set**	**Validation set**	**External test set 1**	**External test set 2**	P_1^A	P_2^B
Number of participants	5548	657	522	400	-	-
Number of eyes	10,357	1191	955	719	-	-
Male/Female	2726/2822	321/336	241/281	190/210	0.20	0.53
Age, years	55.3 (13.4)	56.5 (15.5)	55.4 (13.9)	55.6 (13.3)	0.85	0.64
Hypertension, %	1961 (35.3%)	186 (28.3%)	184 (35.2%)	155 (38.8%)	0.97	0.17
Diabetes, %	3156 (56.9%)	285 (43.4%)	123 (23.6%)	113 (15.7%)	<0.0001	<0.0001
Follow-up duration, months	47.8 (15.8)	56.6 (13.9)	51.4 (11.8)	52.9 (13.5)	<0.0001	<0.0001
Mean time between CFPs, months	42.5 (14.3)	51.4 (12.8)	46.0 (13.2)	47.4 (13.0)	-	-
Median time between CFPs, months	38.0	56.6	46.0	47.0	-	-
Eyes with glaucoma incidence, %	112 (1.1%)	17 (1.4%)	19 (2.0%)	10 (1.4%)	<0.0001	0.02
Glaucoma progression prediction	**Training set (POAG patients)**	**Validation set (POAG patients)**	**External test set 1 (POAG patients)**	**External test set 2 (PACG patients)**	P_1^A	P_2^B
Number of participants	1558	217	172	272	-	-
Number of eyes	3003	422	337	513	-	-
Male/Female	639/919	76/141	76/96	117/155	0.42	0.54
Age, years	44.7 (14.3)	42.8 (13.2)	41.7 (15.2)	45.1 (13.9)	0.008	0.71
Intraocular pressure, mmHg	16.4 (3.6)	16.6 (3.2)	16.2 (3.5)	16.6 (3.7)	0.89	0.27
Mean deviation, dB	−2.6 (4.1)	−2.6 (4.8)	−2.2 (3.6)	−2.7 (4.5)	0.12	0.75
Pattern standard deviation, dB	3.9 (3.9)	3.8 (4.0)	3.2 (3.4)	4.2 (4.3)	0.0009	0.24
Mean times of VF tests, months	7.2 (3.6)	6.5 (2.4)	7.1 (3.5)	6.3 (2.3)	0.67	<0.0001
Mean time between VF tests, months	6.0 (2.0)	6.1 (1.4)	6.4 (2.3)	7.6 (1.1)	-	-
Hypertension, %	197 (12.6%)	26 (12.0%)	31 (18.0%)	35 (12.9%)	0.05	0.92
Diabetes, %	8 (0.5%)	0 (0%)	0 (0%)	0 (0%)	0.35	0.24
Follow-up duration, months	41.7 (4.2)	34.8 (5.8)	39.8 (5.9)	38.4 (7.1)	<0.0001	<0.0001
Eyes with glaucoma progression, %	327 (10.9%)	57 (13.5%)	29 (8.6%)	31 (6.0%)	0.20	0.80

VF, visual field; POAG, primary open-angle glaucoma; PACG, primary angle-closure glaucoma. [A]Comparison of the demographic parameters between training and external test data set 1 by independent *t* test (age, follow-up duration, intraocular pressure, mean deviation, pattern standard deviation, and times of visual field tests) or χ^2 test (sex, cases with hypertension, cases with diabetes, cases with glaucoma diagnosis/incidence/progression). [B]Comparison of the demographic parameters between training and external test data set 2 by independent *t* test (age, follow-up duration, intraocular pressure, mean deviation, pattern standard deviation, and times of visual field tests) or χ^2 test (sex, cases with hypertension, cases with diabetes, cases with glaucoma diagnosis/incidence/progression). All numbers within parentheses are SD.

and a specificity of 0.80 (0.44–0.97) in external test set 2 (Table 2, Figure 2, and Supplemental Figure 3).

Supplemental Table 2 shows the incidence of glaucoma stratified by the AI model. As shown in Supplemental Table 2, there was a significant difference in the incidence rate of glaucoma between the low-risk and high-risk groups. The incidence rates were 0.2% and 5.0%, 0.6% and 5.6%, and 0.4% and 4.1% in the low- and high-risk groups of the validation set, external test set 1, and external test set 2, respectively. We employed the Kaplan-Meier approach to stratify healthy individuals into 2 risk categories (low or high risk) for developing glaucoma, based on 4- to 5-year longitudinal data on glaucoma development. The upper quartile of the predicted risk scores from the model in the validation set was used to create the threshold for the high-risk and low-risk groups in the Kaplan-Meier curves and log-rank tests. In the external test sets, significant separations of the low- and high-risk groups were achieved (both $P < 0.001$, Supplemental Figure 4).

The distribution of the risk scores and the threshold (upper quartile) of low- and high-risk groups across the validation and external test sets are presented in Supplemental Figure 5. As shown in the figure, the threshold (risk score of 0.3561, black dotted line) well defines a boundary to separate individuals who are likely and unlikely to develop glaucoma in a 4- to 5-year period.

Supplemental Table 3 presents the results of subgroup analyses within the validation and external test sets. The AI model demonstrated no statistically significant difference in performance among the subgroups as stratified by age (≥60 vs. <60 years), sex (male vs. female), and severity of glaucoma (mean deviation > −6 dB vs. < −6 dB).

Prediction of the glaucoma progression using longitudinal cohorts. We investigated the predictive performance of the AI model for

Table 2. Performance of the deep-learning models in the validation and external test sets

Glaucoma diagnosis	AUROC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Precision (95% CI)
Validation set	0.97 (0.96–0.97)	0.98 (0.97–0.99)	0.82 (0.80–0.83)	0.99 (0.98–0.99)
External test set 1	0.94 (0.93–0.94)	0.89 (0.87–0.90)	0.83 (0.81–0.84)	0.94 (0.93–0.94)
External test set 2	0.91 (0.89–0.93)	0.92 (0.88–0.96)	0.71 (0.67–0.74)	0.97 (0.95–0.99)
Glaucoma incidence prediction	**AUROC (95% CI)**	**Sensitivity (95% CI)**	**Specificity (95% CI)**	**Precision (95% CI)**
Validation set	0.90 (0.81–0.99)	0.84 (0.82–0.87)	0.82 (0.57–0.96)	1.00 (0.99–1.00)
External test set 1	0.89 (0.83–0.95)	0.84 (0.81–0.86)	0.68 (0.43–0.87)	0.99 (0.98–1.00)
External test set 2	0.88 (0.79–0.97)	0.84 (0.81–0.86)	0.80 (0.44–0.97)	1.00 (0.99–1.00)
Glaucoma progression prediction	**AUROC (95% CI)**	**Sensitivity (95% CI)**	**Specificity (95% CI)**	**Precision (95% CI)**
Validation set	0.91 (0.88–0.94)	0.83 (0.79–0.87)	0.79 (0.66–0.89)	0.96 (0.93–0.98)
External test set 1	0.87 (0.81–0.92)	0.82 (0.78–0.87)	0.59 (0.39–0.76)	0.95 (0.92–0.98)
External test set 2	0.88 (0.83–0.94)	0.81 (0.77–0.84)	0.74 (0.55–0.88)	0.98 (0.96–0.99)

glaucoma progression in glaucomatous eyes over a 3- to 4-year period. A total of 444 POAG eyes had progression within the 3- to 4-year period. The AI model achieved an AUROC of 0.91 (0.88–0.94), a sensitivity of 0.83 (0.79–0.87), and a specificity of 0.79 (0.66–0.89) for predicting glaucoma progression in the validation set (Table 2 and Figure 3). To validate the generalizability of the AI model in predicting progression in multiple-mechanism glaucoma, we further tested its predictive performance in 2 independent cohorts of PACG (external test set 1) and POAG (external test set 2). The AI model achieved excellent predictive performance, with an AUROC of 0.87 (0.81–0.92), a sensitivity of 0.82 (0.78–0.87), and a specificity of 0.59 (0.39–0.76) in external test set 1, and an AUROC of 0.88 (0.83–0.94), a sensitivity of 0.81 (0.77–0.84), and a specificity of 0.74 (0.55–0.88) in external test set 2 (Table 2, Figure 3, and Supplemental Figure 6).

We also trained a predictive model using baseline clinical metadata (age, sex, intraocular pressure, mean deviation, pattern standard deviation, and hypertension or diabetes status) alone to predict progression, which led to an AUROC of 0.76 (0.70–0.82), 0.73 (0.66–0.79), and 0.44 (0.33–0.54) in the validation set, external test set 1, and external test set 2, respectively (Supplemental Figure 7). The performance of the AI model was significantly better than that of the predictive model based on baseline metadata in the above data sets (all $P < 0.001$).

Supplemental Table 4 shows the risk of glaucoma progression stratified by the AI model. As shown in Supplemental Table 4, there was a significant difference in the proportion of eyes with glaucoma progression in the low-risk and high-risk groups. The incidence rates were 3.8% and 42.4%, 4.5% and 23.9%, and 2.0% and 19.8% in the low and high-risk groups of the validation set, external test set 1, and external test set 2, respectively. We then performed Kaplan-Meier analysis to stratify glaucomatous eyes into 2 risk categories (low or high risk) for glaucoma progression, based on 3- to 4-year longitudinal data on glaucoma progression. The upper quartile of the predicted risk scores from the model in the validation set was used to create the threshold for the high-risk and low-risk groups in the Kaplan-Meier curves and log-rank tests. In the external test sets, significant separations of the low- and high-risk groups were achieved (both $P < 0.001$, Supplemental Figure 8).

The distribution of the risk scores and the threshold (upper quartile) of low- and high-risk groups across the validation and external test sets are presented in Supplemental Figure 9. As shown in the figure, the threshold (risk score of 2.6352, black dotted line) well defines a boundary to separate glaucomatous eyes that are likely and unlikely to progress in a 3- to 4-year period.

Supplemental Table 3 presented the results of the subgroup analysis in the validation and external test sets. The AI model demonstrated no statistical significance in all the subgroups stratified by age (≥60 vs. <60 years), sex (male vs. female), and severity of glaucoma (mean deviation > −6 dB vs. < −6 dB) except the AUROCs of severe and less severe subgroups in the validation set and external test set 1.

Visualization of the evidence for prediction of glaucoma incidence and progression. To improve the interpretability of the AI models and illustrate the key regions for AI-based predictions, we used gradient-weighted class activation mapping (Grad-CAM) to generate the key regions in the CFPs for diagnosing glaucoma and predicting glaucoma incidence and progression. Representative cases and their corresponding saliency maps of DiagnoseNet are presented in Supplemental Figure 10. Representative cases and their corresponding saliency maps are presented in Supplemental Figure 10 (DiagnoseNet) and Figure 4 (PredictNet). The saliency maps suggest that the AI model focused on the optic disc rim and areas along the superior and inferior vascular arcades, which is consistent with the clinical approach whereby nerve fiber loss at the superior or inferior disc rim provides key diagnostic or predictive clues. This would suggest that the AI models are learning clinically relevant knowledge in evaluating glaucoma diagnosis and progression. AI-based predictions also appear to involve the retinal arterioles and venules, thus implicating vascular health as potentially relevant to the etiology of chronic open-angle glaucoma.

Discussion

More than 60 million people in the world suffer from glaucoma, and the number is predicted to increase to 110 million by 2040 (24). Due to its insidious onset and variable progression, diagnosis of glaucoma and monitoring of treatment can be challenging and clinically time consuming. Glaucoma screening is not universal around the world, thus leading to a delayed diagnosis and severe irreversible sight loss. Therefore, there is a high clinical demand for an efficient and reliable AI model to help identify high-risk individuals for glaucoma development and progression within the population in order to facilitate early intervention.

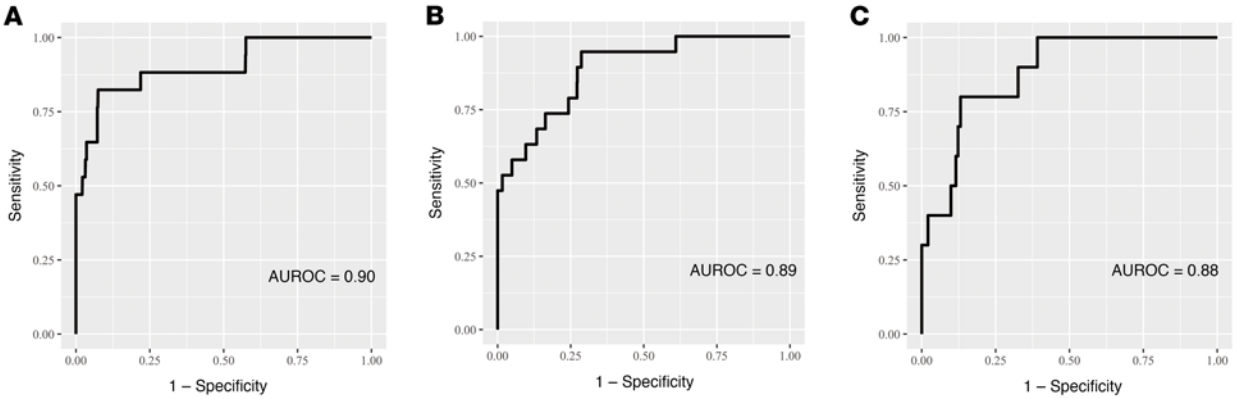


Figure 2. Area under the receiver operating characteristic (AUROC) curves of the AI model for prediction of glaucoma onset. (**A**–**C**) Predictive performance of the AI model in the validation set (*n* = 1191), external test set 1 (*n* = 955), and external test set 2 (*n* = 719).

Deep-learning algorithms have been widely used in glaucoma diagnostic studies (16–19), and have achieved outstanding diagnostic performance in detecting glaucomatous eyes. However, few studies have explored the efficacy of deep learning in glaucoma onset and progression prediction (25–29). In this study, our AI model showed excellent glaucoma diagnostic performance on CFPs, including photographs captured with smartphone cameras using an adaptor, which could significantly broaden its application at a point-of-care setting. Compared with traditional statistical models (30–33), such as glaucoma probability score and Moorfields regression analysis, several studies using deep-learning models achieved comparable or even better predictive performance (25–27). Thakur et al. developed AI models to predict glaucoma development approximately 1 to 3 years before clinical onset and achieved a highest AUROC of 0.88 (25). However, these deep-learning models had some limitations. First, the application was limited to onset prediction without progression prediction, the latter being an essential part of glaucoma management. Second, the data mostly came from hospitals or clinical trials rather than community populations, including many eyes that were diagnosed with ocular hypertension (elevated intraocular pressure without optic neuropathy) rather than glaucoma (25). Third, there is a lack of external validation data to demonstrate the generalizability of the model in the community.

Compared with previous studies, our study has the following advantages. First, we developed AI models for glaucoma diagnosis and incidence and progression prediction. In the external test sets, the models achieved excellent predictive performance in identifying high-risk individuals for developing glaucoma or having glaucoma progression. Secondly, data in glaucoma incidence prediction came from community screening settings, which better reflects the distribution characteristics of glaucoma in the population and facilitates the generalizability of the model. The results in the external data sets show the AI model achieved an excellent predictive performance of glaucoma development, demonstrating strong generalizability and reliability of the AI model. Third, all the patients in the glaucoma cohorts of the progression prediction task have received IOP-lowering medications since enrollment and their IOP values were all controlled within a normal range. This indicates that our predictive model could identify high-risk patients who will undergo glaucoma progression even with reasonably controlled IOPs and facilitate timely interventions such as antiglaucoma surgeries to save vision. Fourth, the AI model based on structural data from CFPs achieved a high predictive accuracy of glaucoma progression, as determined by the gold standard of visual field test results. Visual field tests can reveal functional damage of the optic nerve and are the clinical gold standard in monitoring glaucoma progression (34). As demonstrated in the task of glaucoma progression prediction, the AI model succeeded in identifying the high-risk eyes of progressive functional deterioration from baseline CFPs with high sensitivities. In addition, the AI model showed a similar predictive performance in different subtypes of glaucoma, including POAG and PACG, which share similar structural and functional damage of the optic nerve.

Our study has the following limitations. First, the input data of our AI models are only CFPs. Clinical glaucoma evaluation generally requires integrated analysis of multiple modalities (e.g., clinical examination, optic nerve head imaging, and visual field testing) to determine the glaucoma subtypes and any progression. Our study chose CPFs as the only input due to their high feasibility and widespread availability. Future studies may consider incorporating other data modalities to further improve the predictive performance of the algorithms. Second, only high-quality CFPs were included in the study, which limits the application of the AI models in eyes with media opacities that prevent obtaining clear CFPs. Third, limited by the prevalence of glaucoma in the general population (around 1% to 1.5% in those 40 to 65 years old) (35), there was a relatively small number of cases of glaucoma. To address this issue, we used a deep-learning model with relatively few parameters. Fourth, the AI models presented varied sensitivity and specificity across the data sets, although they had high AUROC values. High sensitivity is more important for screening, and we may further improve the predictive performance of the AI models with more training data in the future. Fifth, all the data were from the Chinese population and further validation is needed in other populations.

In conclusion, our study demonstrates the feasibility of deep-learning systems for disease onset and progression prediction. It offers the possibility of building a virtual glaucoma screening system in the future.

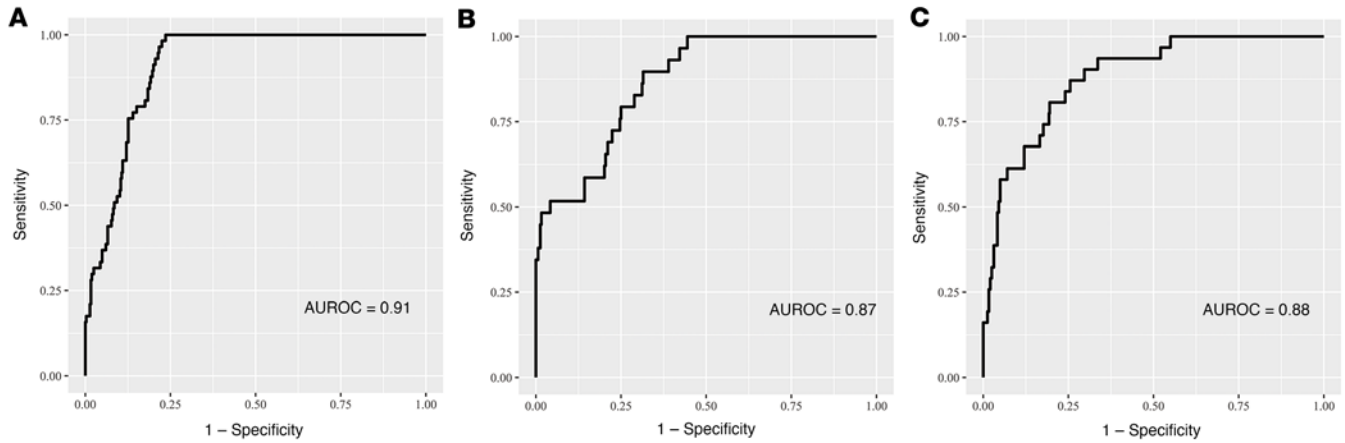


Figure 3. Area under the receiver operating characteristic (AUROC) curves of the AI model for prediction of glaucoma progression. (**A**–**C**) Predictive performance of the AI model in the validation set (*n* = 422), external test set 1 (*n* = 337), and external test set 2 (*n* = 513).

Methods

Data set characteristics

Glaucoma diagnosis cohorts. In these initial cohorts, we were specifically looking for patients visiting ophthalmologists who subspecialize in both glaucoma and anterior segment diseases. The population of patients seen by these ophthalmologists was highly enriched with POAG patients (36, 37). We purposely chose these initial cohorts to ensure that we were able to collect sufficient POAG patients as well as nonglaucomatous control patients (such as cataract patients) who were otherwise appropriately matched for developing an AI-based diagnosis of POAG (Table 1). The training and validation data in glaucoma diagnosis were collected from community cohorts and eye clinics in Guangzhou. To test the generalizability of the AI model, 2 independent data sets obtained from Beijing and Kashi were used as external test sets. The external test set 1 was collected from patients who underwent an annual health check in Beijing city, while the external test set 2 was obtained by smartphones from local eye clinics in Kashi in the Xinjiang Autonomous Region.

Glaucoma incidence prediction cohorts. The training and validation data in the prediction of glaucoma incidence were collected from community cohorts in Guangzhou. To test the generalizability of the AI model, 2 independent data sets obtained from Beijing and Guangzhou communities were used as external test sets. Our longitudinal cohorts for POAG incident prediction had POAG frequencies of around 1% to 2%, which is well within the norm of the prevalence of POAG in the general population.

Glaucoma progression prediction cohorts. The training and validation data in predicting glaucoma progression were collected from 1 POAG cohort in the Zhongshan Ophthalmic Center, Guangzhou. To test the generalizability of the AI model, 2 independent cohorts composed of PACG and POAG eyes from the Zhongshan Ophthalmic Center were used as external test sets.

Image quality control and labeling

Supplemental Figure 11 describes the data sets used in this study and the process of image quality control. All of the images were first deidentified to remove any patient-related information. Fifteen ophthalmologists with at least 10 years of clinical experience were recruited to label the CFPs. First, they were asked to exclude the images with poor quality. The criteria include (a) optic disc or macula was not fully visible and (b) blurred images due to refractive media. A fraction of the CFPs (7.1%) was excluded due to poor quality. Second, the graders were asked to assign glaucoma or nonglaucoma labels to each CFP. Third, each glaucomatous eye with longitudinal follow-up data was further analyzed to determine whether there was a progression based on the visual field reports during follow-up visits. Visual fields with fixation loss lower than 20%, a false positive rate lower than 15%, and a false negative rate lower than 33% were included. Each CFP or visual field report was evaluated by 3 ophthalmologists independently and the ground truths were determined by the consensus of 3 ophthalmologists.

Criteria of glaucoma diagnosis and progression

Glaucoma was diagnosed using the criteria in previous population-based studies (20–22). Glaucomatous optic neuropathy was defined as the presence of vertical cup-to-disc ratio of 0.7 or greater, RNFL defect, optic disc rim width of 0.1-disc diameter or smaller, and/or disc hemorrhage. An eye would be labeled as possible glaucoma if one of the above criteria was met.

Glaucoma progression was determined based on the changes in the visual fields (23). The Humphrey Field Analyzer was used to perform all the visual field tests in 24-2 standard mode (Carl Zeiss Meditec). At least 3 visual field locations worse than baseline at the 5% level in 2 consecutive reliable visual fields, or at least 3 visual field locations worse than baseline at the 5% level in 2 consecutive reliable visual fields, were considered as progression (23). The time of progression was defined as the time from baseline to the first visual field that confirmed progression. Three ophthalmologists examined each visual field report separately to determine progression.

Manual segmentation of anatomical structures

We randomly selected 2000 CFPs for manual segmentations of anatomical structures, including optic disc, optic cup, macula, and blood vessels. Two ophthalmologists independently annotated the CFPs at pixel level, and the final standard reference of annotations was determined by the mean of these 2 independent annotations.

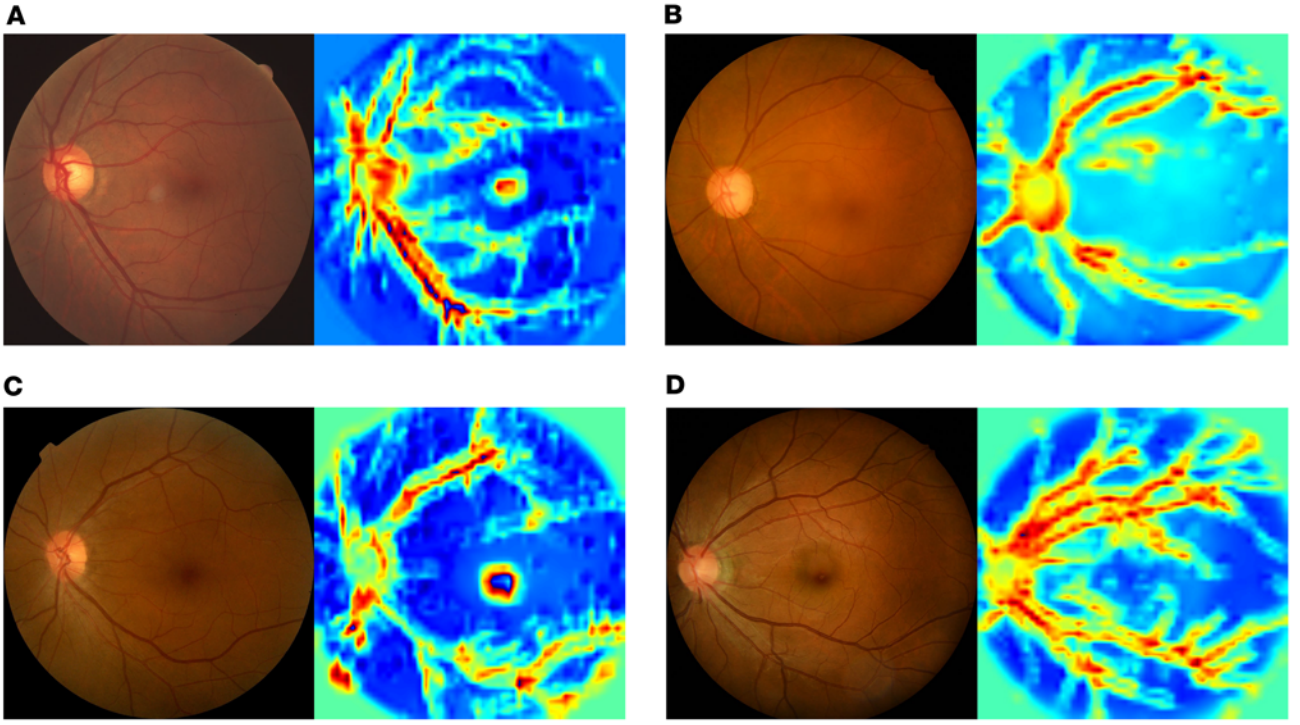


Figure 4. Saliency maps of the deep-learning models. Visual explanation of the key regions the models used for diagnostic predictions. (**A** and **B**) The heatmaps of the typical samples of eyes with (**A**) and without (**B**) glaucoma development. (**C** and **D**) The heatmaps of the typical samples of eyes with (**C**) and without (**D**) glaucoma progression. In both tasks, the saliency maps suggest that the AI model focused on the optic disc rim and areas along the superior and inferior vascular arcades, which are consistent with the clinical approach whereby nerve fiber loss at the superior or inferior disc rim provides key diagnostic clues. AI-based predictions also appear to involve the retinal arterioles and venules.

Model design of glaucoma prediction and ocular disease diagnosis

First, we developed an AI model, DiagnoseNet, to identify CFPs as glaucoma or nonglaucoma. DiagnoseNet is a pipeline composed of modules for segmentation and diagnosis. The fundus images were first semantically segmented in the segmentation module using U-Net (38) to produce 4 anatomical structures: retinal vessels, macula, optic cup, and optic disk. The segmentation data were then merged into a 1-channel by element-wise bit or operation over the 4 anatomical structure–focusing attention layers, which took the place of the CFPs' blue channel to form a new CFP image. The diagnostic module's backbone is EfficientNet-B0, with the last fully connected layer replaced by a Dense layer of 2 output units initialized with a random value, and the other layers' initial weights determined from ImageNet's pretrained settings (Figure 1B).

Then, we created a pipeline, PredictNet, to predict glaucoma onset and progression. PredictNet preprocesses and analyzes the CFP data (Supplemental Figure 1). First, in the preprocessing stage, the original fundus images are enhanced with contrast-limited adaptive histogram equalization (CLAHE) and color normalization (NORM). Important retinal structures, including optic disc, optic cup, macula, and blood vessels are semantically segmented with trained U-Net (38). The multiple-channel anatomical masks generated by U-Net are merged into a 1-channel mask and then fused with the green and red channels of CLAHE images to form CLAHE Normalization Attention–based images. NORM images are fused with the green and red channels of the original images to form anatomical attention–based images. Second, in analyzing stage, CLAHE Normalization Attention–based images and anatomical attention–based images are fed into 2 convolutional neural networks, namely ConvNet-based models 1 and 2. Each ConvNet-based model consists of a feature extraction network and a classification network module. The feature extraction network consists of 3 convolutional blocks, which are composed of a Convolution2D layer, a Batch Normalization layer, a LeakReLu layer, and a MaxPooling2D layer in series, while the classification network consists of 2 Dense layers in series. The GlobalMaxPooling2D layer is used to connect the feature extraction network and the classification network module. The final prediction is obtained by integrating the 2 ConvNet-based models in a linear combination. In the final step, PredictNet will generate a probability (P) of glaucoma incidence or progression between 0 and 1. P was transformed into a z score with the formula z score = $(P - P')/(\text{standard deviation of } P)$, where P' is the mean P of each data set. Then, we obtained the final standard score by adding 1 to all the z scores, because some of the z scores were below zero.

The models were developed with Python (version 3.8.6; https://www.python.org/) and TensorFlow (version 2.1.0; https://github.com/tensorflow/tensorflow). The curves of training loss for each model were generated using TensorBoard (https://github.com/tensorflow/tensorboard) and are presented in Supplemental Figure 12. The key hyperparameters and average running time of each model are summarized in Supplemental Table 5.

Interpretation of the AI model

Grad-CAM (39) was used to highlight the class-discriminative region in the images for predicting the decision of interest. We created heatmaps

generated from CFPs, which indicated the key regions for the AI model to classify the CFPs into low- and high-risk groups.

Data availability

Deidentified data may be available for research purposes from the corresponding authors on reasonable request.

Statistics

The demographic characteristics of study participants are presented as mean ± SD for continuous data, and frequency (percentage) for categorical variables. AUROCs with 95% confidence interval (CI), sensitivity, and specificity were implemented to assess the performance of the algorithms. Sensitivity and specificity were determined by the selected thresholds in the validation sets. The survival curves were constructed for different risk groups, and the significance of differences between groups was tested by log-rank tests. The predictive performance of the AI model and metadata model was performed using DeLong's test. All the hypotheses tested were 2-sided, and a *P* value of less than 0.05 was considered significant. All statistical analyses were performed using R (version 4.0; https://www.r-project.org/).

Study approval

Institutional review board and ethics committee approvals were obtained in all locations and all the participants signed a consent form. All the images were uploaded to a Health Insurance Portability and Accountability Act–compliant cloud server for further grading.

Code availability

The deep-learning models were developed and deployed using standard model libraries and the TensorFlow framework (version 2.3.0). Custom codes were specific to our development environment and used primarily for data input/output and parallelization across computers and graphics processors. The codes are available for research purposes from the corresponding authors on reasonable request.

Author contributions

All authors collected and analyzed the data. KZ and XZ conceived and supervised the project. ZK, ZX, and LF wrote the manuscript. All authors discussed the results and reviewed the manuscript.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC grants 82101117 and 82070955); the High-level Hospital Construction Project, Zhongshan Ophthalmic Center, Sun Yat-sen University (grant 303020104); the Science and Technology Program of Guangzhou, China (2021); the Science and Technology Development Fund (FDCT) of Macau (grant 0070/2020/A2); and FDCT-NSFC (grant 0007/2020/AFJ).

Address correspondence to: Kang Zhang, Macau University of Science and Technology, Avenida Wai Long, Taipa, Macau 999078. Email: kang.zhang@gmail.com. Or to: Xiulan Zhang, Zhongshan Ophthalmic Center, No. 7 Jinsui Road, Guangzhou 510060, China. Email: zhangxl2@mail.sysu.edu.cn.

1. Jonas JB, et al. Glaucoma. *Lancet*. 2017;390(10108):2183–2193.
2. Zhang K, et al. Ophthalmic drug discovery: novel targets and mechanisms for retinal diseases and glaucoma. *Nat Rev Drug Discov*. 2012;11(7):541–559.
3. Weinreb RN, et al. The pathophysiology and treatment of glaucoma: a review. *JAMA*. 2014;311(18):1901–1911.
4. Heijl A, et al. Reduction of intraocular pressure and glaucoma progression: results from the Early Manifest Glaucoma Trial. *Arch Ophthalmol*. 2002;120(10):1268–1279.
5. Jammal AA, et al. Impact of intraocular pressure control on rates of retinal nerve fiber layer loss in a large clinical population. *Ophthalmology*. 2021;128(1):48–57.
6. Gordon MO, et al. Evaluation of a primary open-angle glaucoma prediction model using long-term intraocular pressure variability data: a secondary analysis of 2 randomized clinical trials. *JAMA Ophthalmol*. 2020;138(7):780–788.
7. Medeiros FA, et al. Prediction of functional loss in glaucoma from progressive optic disc damage. *Arch Ophthalmol*. 2009;127(10):1250–1256.
8. Coleman AL, et al. Baseline risk factors for the development of primary open-angle glaucoma in the Ocular Hypertension Treatment Study. *Am J Ophthalmol*. 2004;138(4):684–685.
9. Ocular Hypertension Treatment Study Group, et al. Validated prediction model for the development of primary open-angle glaucoma in individuals with ocular hypertension. *Ophthalmology*. 2007;114(1):10–19.
10. Song Y, et al. Clinical prediction performance of glaucoma progression using a 2-dimensional continuous-time hidden Markov model with structural and functional measurements. *Ophthalmology*. 2018;125(9):1354–1361.
11. Daneshvar R, et al. Prediction of glaucoma progression with structural parameters: comparison of optical coherence tomography and clinical disc parameters. *Am J Ophthalmol*. 2019;208:19–29.
12. De Moraes CG, et al. A validated risk calculator to assess risk and rate of visual field progression in treated glaucoma patients. *Invest Opthalmol Vis Sci*. 2012;53(6):2702–2707.
13. Zhang K, et al. Deep-learning models for the detection and incidence prediction of chronic kidney disease and type 2 diabetes from retinal fundus images. *Nat Biomed Eng*. 2021;5(6):533–545.
14. Wintergerst MWM, et al. Smartphone-based fundus imaging-where are we now? *Asia Pac J Ophthalmol (Phila)*. 2020;9(4):308–314.
15. Iqbal U. Smartphone fundus photography: a narrative review. *Int J Retina Vitreous*. 2021;7(1):44.
16. Li Z, et al. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. *Ophthalmology*. 2018;125(8):1199–1206.
17. Xiong J, et al. Multimodal machine learning using visual fields and peripapillary circular OCT scans in detection of glaucomatous optic neuropathy. *Ophthalmology*. 2021;129(2):171–180.
18. Li F, et al. Development and clinical deployment of a smartphone-based visual field deep learning system for glaucoma detection. *NPJ Digit Med*. 2020;3(1):123.
19. Li F, et al. Digital gonioscopy based on three-dimensional anterior-segment OCT: an international multicenter study. *Ophthalmology*. 2021;129(1):45–53.
20. Iwase A, et al. The prevalence of primary open-angle glaucoma in Japanese: the Tajimi Study. *Ophthalmology*. 2004;111(9):1641–1648.
21. He M, et al. Prevalence and clinical characteristics of glaucoma in adult Chinese: a population-based study in Liwan District, Guangzhou. *Invest Ophthalmol Vis Sci*. 2006;47(7):2782–2788.
22. Foster PJ, et al. The definition and classification of glaucoma in prevalence surveys. *Br J Ophthalmol*. 2002;86(2):238–242.
23. Garway-Heath DF, et al. Latanoprost for open-angle glaucoma (UKGTS): a randomised, multicentre, placebo-controlled trial. *Lancet*. 2015;385(9975):1295–1304.
24. Tham YC, et al. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology*. 2014;121(11):2081–2090.
25. Thakur A, et al. Predicting glaucoma before onset using deep learning. *Ophthalmol Glaucoma*. 2020;3(4):262–268.
26. Thakur A, et al. Convex representations using deep archetypal analysis for predicting glaucoma. *IEEE J Transl Eng Health Med*. 2020;8:1–7.
27. Gupta K, et al. Glaucoma precognition based on confocal scanning laser ophthalmoscopy images of the optic disc using convolutional neural network. 2021 IEEE/CVF Conference on

Computer Vision and Pattern Recognition Workshops (CVPRW). 2021;2259–2267. https://doi.org/10.1109/CVPRW53098.2021.00255.
28. Shuldiner SR, et al. Predicting eyes at risk for rapid glaucoma progression based on an initial visual field test using machine learning. *PLoS One*. 2021;16(4):e0249856.
29. Wen JC, et al. Forecasting future Humphrey Visual Fields using deep learning. *PLoS One*. 2019;14(4):e0214875.
30. Zangwill LM, et al. Baseline topographic optic disc measurements are associated with the development of primary open-angle glaucoma: the Confocal Scanning Laser Ophthalmoscopy Ancillary Study to the Ocular Hypertension Treatment Study. *Arch Ophthalmol*. 2005;123(9):1188–1197.
31. Weinreb RN, et al. Predicting the onset of glaucoma: the confocal scanning laser ophthalmoscopy ancillary Study to the Ocular Hypertension Treatment Study. *Ophthalmology*. 2010;117(9):1674–1683.
32. Gordon MO, et al. The Ocular Hypertension Treatment Study: baseline factors that predict the onset of primary open-angle glaucoma. *Arch Ophthalmol*. 2002;120(6):714–720.
33. Salvetat ML, et al. Baseline factors predicting the risk of conversion from ocular hypertension to primary open-angle glaucoma during a 10-year follow-up. *Eye (Lond)*. 2016;30(6):784–795.
34. Hu R, et al. Functional assessment of glaucoma: uncovering progression. *Surv Ophthalmol*. 2020;65(6):639–661.
35. Rudnicka AR, et al. Variations in primary open-angle glaucoma prevalence by age, gender, and race: a Bayesian meta-analysis. *Invest Ophthalmol Vis Sci*. 2006;47(10):4254–4261.
36. Li J. Glaucoma type proportion of glaucoma outpatient in Beijing Tongren Hospital from 2014 to 2016. *Investig Opthalmol Vis Sci*. 2018;59(9):2745.
37. Song YJ, et al. Comparison of glaucoma patients referred by glaucoma screening versus referral from primary eye clinic. *PLoS One*. 2019;14(1):e0210582.
38. Ronneberger O, Fischer P, and Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, et al, eds. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer International Publishing; 2015(3):234–241.
39. Selvaraju RR, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. Paper presented at: 2017 IEEE International Conference on Computer Vision (ICCV); October 22–29, 2017; Venice, Italy. https://doi.ieeecomputersociety.org/10.1109/ICCV.2017.74. Accessed April 14, 2022.