tvst

Artificial Intelligence

A Multicenter Clinical Study of the Automated Fundus Screening Algorithm

Fei Li[1,*], Jianying Pan[1,*], Dalu Yang[2,*], Junde Wu[3,*], Yiling Ou[1], Huiting Li[1], Jiamin Huang[1], Huirui Xie[1], Dongmei Ou[1], Xiaoyi Wu[1], Binghong Wu[2], Qinpei Sun[2], Huihui Fang[2], Yehui Yang[2], Yanwu Xu[2], Yan Luo[1], 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, Guangzhou, China
[2] Intelligent Healthcare Unit, Baidu, Beijing, China
[3] Kangfuzi, Beijing, China

Correspondence: Yanwu Xu, Baidu Inc., Beijing 100085, China.
e-mail: ywxu@ieee.org
Yan Luo, Zhongshan Ophthalmic Center, Guang Zhou 510060, Guang Dong, China.
e-mail: luoyan2@mail.sysu.edu.cn
Xiulan Zhang, Zhongshan Ophthalmic Center, Guang Zhou 510060, Guang Dong, China.
e-mail: zhangxl2@sysu.edu.cn

Received: March 8, 2022
Accepted: June 16, 2022
Published: July 26, 2022

Keywords: automated fundus screening software; diabetic retinopathy; macular dystrophies; glaucoma

Citation: Li F, Pan J, Yang D, Wu J, Ou Y, Li H, Huang J, Xie H, Ou D, Wu X, Wu B, Sun Q, Fang H, Yang Y, Xu Y, Luo Y, Zhang X. A multicenter clinical study of the automated fundus screening algorithm. Transl Vis Sci Technol. 2022;11(7):22, https://doi.org/10.1167/tvst.11.7.22

Purpose: To evaluate the effectiveness of automated fundus screening software in detecting eye diseases by comparing the reported results against those given by human experts.

Results: There were 1585 subjects who completed the procedure and yielded qualified images. The prevalence of referable diabetic retinopathy (RDR), glaucoma suspect (GCS), and referable macular diseases (RMD) were 20.4%, 23.2%, and 49.0%, respectively. The overall sensitivity values for RDR, GCS, and RMD diagnosis are 0.948 (95% confidence interval [CI], 0.918–0.967), 0.891 (95% CI, 0.855–0.919), and 0.901 (95% CI–0.878, 0.920), respectively. The overall specificity values for RDR, GCS, and RMD diagnosis are 0.954 (95% CI, 0.915–0.965), 0.993 (95% CI–0.986, 0.996), and 0.955 (95% CI–0.939, 0.968), respectively.

Methods: We prospectively enrolled 1743 subjects at seven hospitals throughout China. At each hospital, an operator records the subjects' information, takes fundus images, and submits the images to the Image Reading Center of Zhongshan Ophthalmic Center, Sun Yat-Sen University (IRC). The IRC grades the images according to the study protocol. Meanwhile, these images will also be automatically screened by the artificial intelligence algorithm. Then, the analysis results of automated screening algorithm are compared against the grading results of IRC. The end point goals are lower bounds of 95% CI of sensitivity values that are greater than 0.85 for all three target diseases, and lower bounds of 95% CI of specificity values that are greater than 0.90 for RDR and 0.85 for GCS and RMD.

Conclusions: Automated fundus screening software demonstrated a high sensitivity and specificity in detecting RDR, GCS, and RMD from color fundus imaged captured using various cameras.

Translational Relevance: These findings suggest that automated software can improve the screening effectiveness for eye diseases, especially in a primary care context, where experienced ophthalmologists are scarce.

Introduction

Diabetic retinopathy (DR), glaucoma, and various kinds of macular degeneration (MD) are the leading causes of blindness.[1] According to Teo et al.,[2] 28.54 million adults worldwide have vision-threatening DR in 2020 and the number can increase to 44.82 million in 2045. For glaucoma, the worldwide prevalence is projected as 76.0 million for 2020 and 111.8 million in 2040.[3] The two most common types of MD are age-related MD and myopic MD. In 2020, the prevalence of age-related MD is approximately 196 million and can increase to 288 million in 2040, and in 2015 myopic MD has a prevalence of 10.0 million and is projected to 55.7 million in 2050.[4,5]

If these eye diseases can be detected in their early stage, the vision loss is able to be slowed or prevented. However, the early detection of these diseases requires the examinations of high-level ophthalmologists, which are usually not available for large-scale screening. Deep learning (DL) has been widely applied in automatic fundus photograph analysis, especially in detecting vision-threatening eye diseases.[6] Various algorithms and clinical trials have been developed and conducted for detecting DR,[7–10] glaucoma suspects (GCS),[11–13] and MD.[14–17] Recent studies also showed these algorithms are able to predict multiple diseases simultaneously. Ting et al.[18] used a DL system to perform detection of RDR, vision-threatening DR, GCS, and RMD at the same time in a multiethnic population. Islam et al.[19] applied a DL method to predict six diseases, namely, RDR, RMD, GCS, cataract, hypertensive retinopathy, and myopia.

Here we present a prospective study to validate the performance of an artificial intelligence (AI) screening system for the detection of multiple eye diseases. We validate that the AI screening system can accurately and robustly detect referable DR (RDR), referable macular diseases (RMD), and GCS among fundus images of various cameras. We present detailed comparison of sensitivity, specificity, and area under the curve (AUC) values across different domains of interest. Our main hypotheses are that lower bounds of the 95% confidence interval (CI) of sensitivity values are greater than 0.85 for all three target diseases, and the lower bounds of the 95% CI of specificity values are greater than 0.90 for RDR and 0.85 for GCS and RMD.

Methods

AI Fundus Screening Algorithm

The automated screening algorithm consists of three major modules, namely, a structural analysis, quality assessment, and disease prediction. In the structural analysis module, we perform optic disc detection and fovea detection to locate the region of interest (ROI) that is relevant to GCS and RMD, respectively. The quality assessment module decides whether each ROI (as well as the whole image) is of low quality based on a comprehensive analysis of the brightness, contrast, and blurriness. If any of the ROIs is determined as low quality, the image will be disqualified from the study and the system will prompt a retake. In the disease prediction module, we use the self-designed DL models to predict the diseases of interest based on the given images. The disease prediction module consists of three self-designed DL models, which are designed to predict RDR, RMD and GCS. The backbones of the three models are designed by similar architectures.

The backbone of the disease prediction module is a combination of DenseNet-121[23] and bilinear pooling. DenseNet-121 is a widely used general network structure proposed in recent years. The structure alleviates the gradient disappearance phenomenon during model training by stacking multiple dense blocks with connections, and strengthens the internal feature reuse of the model, so as to achieve a high accuracy of the model. In addition, to obtain the better performance, following unique strategies are adopted:

a) The input image size is increased from 224 × 224 to 512 × 512, which helps to better preserve the detail information of the original image.
b) The bilinear pooling layer is used to replace the gap layer in the original DenseNet-121. As a general technique in the field of fine-grained image classification, a bilinear pooling layer can help the model to extract the texture information in the image (such as DR-related bleeding/optic disc and cup structure), so as to help model focus on discriminative features.
c) In terms of DR detection, the model will output fine grained grading results (no DR/phase I/phase II/phase III/phase IV or above), and then transform it into binary or ternary classification results through probability weighting. When modeling and optimizing the model, an automated screening algorithm encodes the classification label by ordinal regression. Compared with one-hot labels, ordinal regression labels impose stronger penalty when the deviation is larger, so as to facilitate the grading.

In addition, to better distinguish patients with other ophthalmic diseases from the target diseases, the algorithm adds an additional “abnormal” category in the model to decrease the intraclass divergence during training

In this study, the automated screening algorithm system runs locally on a desktop computer (Intel i5-8400 and NVIDIA GeForce GTX 1060). The participants will receive their AI screening reports within 30 seconds after taking the images. An AI screening report shows whether participants were positive for RDR, RMD, and GCS. The predicted possibilities of each disease are also recorded internally to calculate the AUC. The results are compared against the labels given by ophthalmologists for evaluating the performance.

Table 1. Internal Validation Performance of the AI Screening Algorithm

Target Disease	Sensitivity (95% CI)	Specificity (95% CI)
RDR	0.944 (0.914–0.967)	0.977 (0.965–0.986)
GCS	0.965 (0.947–0.980)	0.938 (0.917–0.954)
RMD	0.913 (0.878–0.942)	0.910 (0.890–0.928)

Before deployment of the algorithm, an internal validation had been conducted to verify the effectiveness. The experiments are conducted on an internal dataset with 1229 scans, which contains 327 positive cases of RDR, 324 positive cases of GCS, and 527 positive cases of RMD. The internal validation results are shown in Table 1.

Study Design

We prospectively enrolled the subjects at seven hospitals throughout China. Patients who visited the designated hospitals between April 16, 2020, and October 21, 2020, who were more than 18 years old, and who were able to cooperate with fundus photography were invited to participate. We excluded patients who previously underwent eye surgery, were pregnant, or were with ocular media opacification, because these cases may affect the quality of fundus photos.

We obtained approval for the study protocol from ethical committees of all the seven hospitals (the clinical trial registration information can be checked at https://github.com/BaiduIHU/Clinical-trial-registration-AI100/tree/main). We ensured that the subjects signed written informed consent and were over 18 years old. The study was funded by Kangfuzi Inc. and designed by the authors. Xiangkang Inc., provided data management and lockdown, equipment maintenance, monitoring, and auditing services at all hospitals, as a contract research organization.

Study Protocol

The study protocol consists of the following steps:

1. A participant signs written informed consent.
2. An operator records a participant's demographic information and medical history, then confirms whether to enroll the participant according to the selection criteria of this study. The inclusion criteria are as discussed elsewhere in this article.
3. The operator encodes the participant's information according to authentication code table after enrollment, then takes images with a fundus camera according to the standard imaging protocol. One posterior pole image is taken for each eye.
4. The operator submits the images to the IRC. The IRC determines whether the images are readable.
5. For the readable images, the operator uploads the images to automated screening algorithm for the automatic analysis. The operator then prints out, confirms, and signs on the automatic analysis results. Meanwhile, the IRC grades the images according to the study protocol and documents the grading results.

Grading Standards

Overall Grading Procedure

We invited two certified ophthalmologists in IRC as the graders. The two graders had 6 and 10 years of clinical experience. Each grader reads the fundus images according to the following standard procedure.

1. Image quality assessment. The grader first assesses whether the fundus image is readable or not. If the image is not readable, the grader performs no further grading of diseases and records the final result as low quality. Subjects with low-quality images are not counted for final performance calculation. The criteria of image quality assessment are described elsewhere in this article.
2. Optic disc, optic cup, and macula boundary determination. For a readable image, the grader first determines the boundary of the optic disc and optic cup, and roughly estimates their respective diameters. The grader then locates the fovea and determines the macula boundary as a circle centered at the fovea with a certain radius, which is the minimum of the three lengths: (i) two times the optic disc diameter; (ii) the minimum distance between the fovea and the optic disc boundary; and (iii) the minimum distance from the fovea to the two main branches of the central retinal vein.
3. Grading of RDR, RMD, and GCS. The grader then examines the referable diseases in the images. For the grading of RMD, the grader uses the macula boundary determined in step 2 as the ROI. For the grading of GCS, the grader considers the cup-disc-ratio based on the estimated optic disc and optic cup diameters. The detailed reference standards of the diseases are shown in the later sections. Subjects who have two or three diseases will be separately counted in each class. We classified each image as positive or negative

for GCS, RDR, and RMD. We do not pay special consideration when analyzing the images with two or more diseases.

4. For a fundus image, two graders independently examine the image and give their first opinion of the three diseases. If any of the first opinions differ, the result are arbitrated by a senior ophthalmologist with 19 years of clinical experience. The arbitrated results are documented on the IRC grade recording forms as the final grading results.
5. We ensured that all the graders were masked to all patient information, including age, sex, medical history, previous diagnosis, and other clinical test results.

Quality Assessment Criteria

The grader will disqualify an image if any of the following occurs.

1. The optic disc is not fully visible in the field of view.
2. One-third or more of the macular region is not visible in the field of view.
3. The optic disc is too bright owing to overexposure.
4. The macular region is too dark owing to underexposure or a small pupil.
5. The image is out of focus.
6. The image is blurred owing to motion, blinking, ocular media opacity, and so on.

Reference Standard for RDR, RMD, and GCS

For RDR, we first follow Ting et al.[6] to determine the DR severity scale. In this study, we categorize a subject as RDR positive if its DR severity level is greater than or equal to II (moderate nonproliferative). A typical fundus image with RDR contains at least one microaneurysm and at least one hemorrhage or exudate. It is worth noting that this study follows the China RDR classification standard. In China's classification standard, grades I, II, and III correspond with grades I, II, and III of international standard, respectively; grades IV, V, and VI correspond with grade IV of international standard. This discrepancy will not affect the conclusions of the study.

For GCS, we follow the suspected glaucoma hallmarks listed in Zhang et al.,[22] including an enlarged disc cup ratio, optic disc pallor, rim widths that do not fit the inferior ≥ superior ≥ nasal ≥ temporal rule, and so on. If the fundus image fits two or more of the suspected glaucoma hallmarks, the subject is determined as GCS positive.

In community screening settings, various macular abnormalities are prevalent. However, without further examinations other than fundus photography, it is difficult to make a differential diagnosis among all types of macular diseases owing to their complicated and nonspecific appearances. Thus, we combined a referable macular lesion list in Zhang et al.,[22] and automated screening algorithm was developed to detect the set of macular lesions. The referable lesion list includes drusen (of diameter >125 µm), exudate, hemorrhage, atrophy, epiretinal membrane, macular hole, retinal detachment, pigment epithelial proliferation, and so on. If the fundus image contains any of the referable lesions in the macular area, the subject is determined as RMD positive.

Statistical Analyses

The primary effectiveness evaluation criteria are sensitivity and specificity. The statistical tests of sensitivity and specificity are two sided. Only predictions that are greater than the predetermined threshold are counted in the analysis. We consider the AI algorithm meets the clinical usage requirement when the lower bounds of the 95% CIs exceed the predetermined end points.

The sample size of the study is determined by the significance level, the power of test, and the estimated and target values of sensitivity and specificity. The formula for calculating the sample size (for each target disease) is as follows:

$$N = \frac{\left[Z_{1-\alpha}\sqrt{P_0\left(1-P_0\right)} + Z_{1-\beta}\sqrt{P_T\left(1-P_T\right)}\right]^2}{\left(P_T - P_0\right)^2},$$

where $\alpha = 0.05$ is the level of significance, $1 - \beta = 8$ is the power of the test, P_0 is the least acceptable value for the sensitivity (or specificity) of the clinical study, and P_T is the estimated sensitivity (or specificity) set in reference to the internal validation results. The calculated sample sizes needed for RDR, GCS, and RMD are shown in Table 2. Because it is possible to share the negative samples, the total estimated sample size was 1557 (the sum of 282, 363, 363, and 549). Considering withdrawal and exclusion rates of approximately 10%, we determine the total target sample size as 1730.

The sensitivity and specificity are computed with the following formulas:

$$\mathrm{Se} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},\ \mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}$$

where TP is a true positive, TN a true negative, FP a false positive, and FN a false negative (Table 3). The 95% CIs of sensitivity and specificity are calculated

Table 2. Estimated Sample Size Needed for Each Target Disease

	P_T Sensitivity	P_0 Sensitivity	P_T Specificity	P_0 Specificity	Calculated *N* Positive	Calculate *N* Negative
RDR	0.9	0.85	0.93	0.9	282	549
GCS	0.9	0.85	0.9	0.85	363	363
RMD	0.9	0.85	0.9	0.85	363	363

Table 3. Contingency Table Used for Sensitivity and Specificity Calculation

IRCAlgorithm	Positive	Negative
Positive	TP	FP
Negative	FN	TN

using the Wilson method. All statistical analyses are performed with the use of SAS software, version 9.3 (SAS Institute, Cary, NC).

Results

Study Population

In this study we initially selected a total of 1743 subjects. There were 1738 subjects who completed the procedures and had fundus images taken. We excluded 153 subjects in the final full analysis set because they quit, withdrew informed consent, did not meet the inclusion criteria, repeatedly enrolled, or had a recorded fundus images with low quality. As a result, the full analysis set contains 1585 clear fundus images from 1585 subjects (1 eye from each subject). The average age of the subjects was 53.19 ± 15.59. Of the 1585 subjects, 900 were male. Detailed demographics of the population are presented in Table 4. The prevalence of RDR, GCS, and RMD were 20.4% (334/1585), 23.2% (368/1585), and 49.0% (777/1585), respectively. Further, for the 322 RDR patients, 10 of them have mild nonproliferative DR, 229 of them have moderate nonproliferative DR, 60 of them have severe nonproliferative DR; the rest 35 have PDR. Detailed demographics of the disease prevalence are presented in Table 5.

Table 4. Demographic of the Population

Gender, *n* (%)	
Total	1585 (100.00)
Male	900 (56.78)
Female	685 (43.22)
Age (years)	
Mean (SD)	53.19 ± 15.59
Median (Q1, Q3)	56.00 (43.00, 65.00)
Minimum, maximum	18.00, 91.00
Age frequency distribution, *n* (%)	
Total	1585 (100.00)
<20	10 (0.63)
20–29	161 (10.16)
30–39	159 (10.03)
40–49	238 (15.02)
50–59	387 (24.42)
60–69	410 (25.87)
≥70	220 (13.88)
Ethnic group, *n* (%)	
Total	1585 (100.00)
Han Chinese	1530 (96.65)
Others	53 (3.35)
Unknown	2 (0.13)

Table 5. Demographics of Disease Prevalence

Clinical Diagnosis	*n* (%)
Normal	537 (33.88%)
RDR only	12 (0.76%)
GCS only	255 (16.09%)
RMD only	364 (22.97%)
RDR and RMD	304 (19.18%)
RDR and GCS	4 (0.25%)
GCS and RMD	95 (5.99%)
RDR and RMD and GCS	14 (0.88%)
Total	1585 (100%)

Screening Performance of the Automated Screening Algorithm

We evaluated the AI screening performance by comparing the diagnosis results of automated screening algorithm against those given by the ophthalmologists. We report an overall evaluation in terms of sensitivity, specificity, AUC, and their respective 95% CIs, across all clinical trial centers and camera models.

Table 6. Overall Performance of AI Screening Algorithm

Target Disease	Sensitivity (95% CI)	Specificity (95% CI)	AUC (95% CI)	Confusion Matrix (TP, FN, FP, TN)
RDR	0.948 (0.918–0.967)	0.954 (0.915–0.965)	0.976 (0.968–0.985)	307, 17, 57, 1204
GCS	0.891 (0.855–0.919)	0.993 (0.986–0.996)	0.990 (0.985–0.995)	328, 40, 9, 1208
RMD	0.901 (0.878–0.920)	0.955 (0.939–0.968)	0.968 (0.960–0.976)	700, 77, 36, 772

Table 7. Agreement Between the AI Screening Algorithm and the Two Graders

	AI vs. Grader 1 (6 Years) Intraclass Correlation Coefficient	AI vs. Grader 2 (10 Years) Intraclass Correlation Coefficient
RDR	0.893	0.915
RMD	0.914	0.921
GCS	0.942	0.948

We also show the separate evaluation results of each camera brand.

Overall Performance

Among 1585 subjects, 363, 339, and 737 subjects are diagnosed by the algorithm as RDR, GCS, and RMD, respectively. The overall sensitivity values for RDR, GCS, and RMD are 0.948 (95% CI, 0.918–0.967), 0.891 (95% CI, 0.855–0.919), and 0.901 (95% CI, 0.878–0.920), respectively. The overall specificity values for RDR, GCS, and RMD are 0.954 (95% CI, 0.915–0.965), 0.993 (95% CI–0.986, 0.996), and 0.955 (95% CI, 0.939–0.968), respectively. The overall AUC values for RDR, GCS, and RMD are 0.976 (95% CI, 0.968–0.985), 0.990 (95% CI, 0.985–0.995), and 0.968 (95% CI, 0.960–0.976), respectively. The complete results are shown in Table 6. We also assess the agreement between the AI algorithm and the two graders respectively. The intraclass correlation coefficients are provided in Table 7.

Performance by Camera Brands

In this study, we also validated the automated screening algorithm on three different brands of cameras that are used in the different centers. Detailed information about these cameras is shown in Table 8. For convenience, we use camera I, II, and III to denote Topcon, Syseye, and Zeiss cameras, respectively, in the following analysis. Centers A, C, D, and H use camera brand I. Centers B and F use camera brands II and III, respectively. Center G uses a mixture of camera brands II and III. The total numbers of subjects are 981, 281, and 323 for camera brands I, II, and III, respectively. In Table 9, we report the sensitivity, specificity, AUC, and numbers of positive samples for RDR, GCS, and RMD diagnosed by ophthalmologists.

Screening Performance of Individual Ophthalmologists

As a reference, we also compute the screening performance in terms of sensitivity and specificity of two independent graders. The ground truth of this experiment is the same as which of automated screening algorithm. The result values are shown in Table 10.

Further, we investigated the interobserver agreement of the three target diseases between the two independent graders. The confusion matrix is reported in Table 11.

Table 8. Information About the Three Different Cameras We Used

Brand	Topcon	Syseye	Zeiss
Model	TRC-NW400	RetiCam 3100	VISUCAM 200
Mode	Non Mydriatic	Non Mydriatic	Non Mydriatic
Setting	Desktop, Automatic	Desktop, Automatic	Desktop, Manual
Fixation	Center	Center	Center
Resolution	1956 × 1934	2656 × 1992	2124 × 2056
Minimum pupil size	3.3 mm	2.8 mm	3.3 mm
Field of view	45°	50°	45°

Table 9. Performance of the Automated Screening Algorithm in Detecting Three Diseases, Breakdown by Camera Model

Target Disease	Camera Model	N-Positive	Sensitivity (95% CI)	Specificity (95% CI)	AUC (95% CI)
RDR	I	141	0.950 (0.900–0.980)	0.968 (0.954–0.979)	0.985 (0.977–0.993)
	II	54	0.944 (0.846–0.988)	0.912 (0.867–0.945)	0.944 (0.912–0.976)
	III	129	0.945 (0.890–0.978)	0.944 (0.902, 0.972)	0.982 (0.968–0.995)
GCS	I	262	0.855 (0.806–0.895)	0.996 (0.988–0.999)	0.991 (0.984–0.997)
	II	65	0.969 (0.893–0.996)	0.968 (0.934–0.987)	0.991 (0.982–1.000)
	III	41	1.000 (0.914–1.000)	0.996 (0.980–1.000)	0.999 (0.997–1.000)
RMD	I	403	0.854 (0.815–0.887)	0.971 (0.953–0.983)	0.964 (0.951–0.977)
	II	165	0.970 (0.931–0.990)	0.853 (0.776–0.912)	0.972 (0.953–0.990)
	III	209	0.938 (0.896–0.963)	0.965 (0.913–0.990)	0.981 (0.969–0.993)

Table 10. Independent Grader Results Compared With Arbitrated Ground Truth From IRC

	Grader 1 Sensitivity	Grader 1 Specificity	Grader 2 Sensitivity	Grader 2 Specificity
RDR	0.963 (0.936–0.981)	0.989 (0.981–0.994)	0.985 (0.964–0.995)	0.995 (0.990–0.998)
GCS	0.976 (0.954–0.989)	0.993 (0.986–0.997)	0.989 (0.972–0.997)	0.992 (0.985–0.996)
RMD	0.983 (0.972–0.991)	0.995 (0.987–0.999)	0.988 (0.978–0.995)	0.994 (0.986–0.998)

Table 11. Interobserver Agreement Between the Two Independent Graders (Without Adjudication)

Disease	Grader 2/Grader 1	Positive	Negative	Kappa
RDR	Positive	307	19	0.928
	Negative	18	1241	
GCS	Positive	355	13	0.943
	Negative	19	1198	
RMD	Positive	757	11	0.966
	Negative	16	801	

Discussion

The purpose of this study was to analyze the performance and robustness of automated screening algorithm on fundus images. According to the study design, the end points are the lower bounds of the 95% CI of sensitivity values that are greater than 0.85 for all three target diseases, and the lower bounds of the 95% CI of the specificity values that are greater than 0.90 for RDR and 0.85 for GCS and RMD. The results show that the performance exceed the end points by a fair margin.

In Table 6, we can see the automated screening algorithm detects all three target diseases with sensitivity values that are greater than 0.85, indicating its significant application potential in primary care sites. In general, primary care sites lack professional ophthalmologists. These highly sensitive automated screening algorithm are an important complementary tool for screening in the early stage fundus disease. In addition, the specificity values of the automated screening algorithm are also greater than 0.90 across all three diseases. In a primary care setting with a large-scale population but few positive cases, a high specificity value means fewer false positives, which could help to conserve valuable high-level health care resources. In addition, the short screening time, consistency of the results, and the narrow CI indicate that this AI automated screening device is a reliable screening tool in various population groups.

The overall AUCs for detecting all the three diseases are high, indicating that the automated screening algorithm can correctly distinguish positive features from negative ones in three target diseases. The experimental results also show that the automated screening algorithm has good generalizability to different brands of cameras. In Table 9, we can see the algorithm remains high AUCs (>0.94) for different cameras. The sensitivity and specificity values for all three diseases are greater than 0.85. The corresponding lower bounds of the 95% CIs are greater than 0.80, except for the specificity value of RMD in camera brand II (0.776). This strong generalizability facilitates deployment in primary care sites that cannot afford expensive equipment.

In this study, we followed two standard grading procedures[16] for grading. The interobserver comparison results show good consistency for discrimination between disease presence and absence, which indicates the study conducted a reliable evaluation on the AI algorithm screening performance.

Compared with validations of the other registered automated screening devices, like the IDx-DR,[25] CARE,[20] and Airdoc,[26] this prospective study is the first to include multiple retina pathologies in the evaluation. Moreover, most similar studies evaluate the AI devices on only one brand of camera. However, the AI algorithms are notoriously vulnerable to the domain shift,[27,27] which means an AI device performs well in one center, whereas it may show an unacceptable performance in another center in its application. This performance gap stems from the shift of data distribution, which is usually caused by the camera differences. This study validated the algorithm performance on four different brands of cameras, which shows the fair generalization ability of this automated screening algorithm.

There are also several limitations in this study. First, this study was conducted in hospital settings and recruited subjects from within the ophthalmology department, where the population distribution does not align with community screening or physical examination scenarios. In hospitals, there are more patients with more severe diseases, whereas in general population screening scenarios, there are more healthy people and patients with milder and early stages of the diseases. Thus, these study results cannot directly indicate the performance of the AI software in community screening scenarios. Actually, we have conducted another comprehensive study to compare the AI performance in an in-hospital scenario and community screening scenario. It showed that the evaluation results on general population will have a slight decrease in the sensitivity (approximately 1%–2%), but with an obvious increase in specificity (approximately 3%–6%) in the community screening scenario on all three diseases. Detailed results and a discussion will be provided in our future work.

Second, the ground truth of GCS may not be completely correct. In this study, graders are blind to all other clinical information except the fundus image. However, a diagnosis based on fundus images is not the gold standard for glaucoma. Although the graders have a good agreement on GCS, this agreement may be biased. Thus, the automated screening algorithm may report false results regarding glaucoma, even if it agrees with the graders. In contrast, the validated AI model can only predict suspected glaucoma, but cannot confirm the diagnosis. Clinically, glaucoma is diagnosed through a combination of the visual field test, intraocular pressure measurement, gonioscopy, OCT, and color fundus photography-based optic nerve assessment. Thus, an all-around AI model needs to detect the glaucoma based on multimodal data. To date, there is still no AI model that can combine all these test results for an automated diagnosis. However, several recent studies are working toward this goal. For example, in the GAMMA challenge[24] we organized recently verified well-designed AI models that can detect glaucoma from a combination of fundus images and OCT volumes and achieve a superior performance. A promising way to automatically confirm a glaucoma diagnosis in the future is to further combine intraocular pressure measurement data and visual field test data to create an automated glaucoma diagnosis AI model in full accordance with the clinical glaucoma diagnosis criteria.

Third, in this study, RMD is defined as a combination of several referable lesions. However, in real-world

Table 12. The Examples of False Negatives and False Positives

Target Disease	False Negative	False Positive
RDR	1. Confusion between minor hemorrhages and microaneurysm	1. Confused with hypertensive retinopathy
	2. RDR signs appear at the edge of the image	2. Confused with retinal vein occlusion
	3. Interference of epiretinal membrane	3. Confused with retinitis pigmentosa
GCS	1. Interference of myopic crescent	Nine cases in total, with no obvious patterns
	2. The optic disc is too bright, which results in a smaller cup to disc ratio being detected	
RMD	1. Drusen of critical referable size not detected	1. Camera lens stains
	2. The lesion at the border of the macular area was not counted in the macular area	2. Severe tessellation
	3. Central serous chorioretinopathy not detected	

practice, the types of the referable lesions (such as drusen and pigment epithelial proliferation) depend on local medical guidelines. Sometimes, it also requires long-term observation and comparisons to evaluate the progress of the diseases. Thus, the diversity of reference standards may limit the application scope of automated screening algorithm.

Finally, although the performance of AI algorithm exceeds the end points by a fair margin, there still have several false positives and false negatives in all three diseases. We analyzed the false-positive examples and false-negative examples in detail. The main examples are presented in Table 12.

In Table 11, it can be inferred that most false positives are due to the interference of other diseases or lesions. For example, the prediction of RDR is interfered with hypertension fundus, venous obstruction, retinitis pigmentosa, minor hemorrhages, and so on. One way to solve this may be to train a more comprehensive AI model to discriminate against these diseases and lesions. Another possible solution is to combine some other medical records and examinations to achieve more reliable predictions. Some other false positives are caused by technical problems, such as the stains on the camera lens or a limited field of view. However, these technical pitfalls cannot be avoided in real-world use, especially in primary care.

Acknowledgments

Supported by Kangfuzi Inc.

The datasets generated during and/or analyzed during the current study are not publicly available owing to a business confidentiality agreement, but are available from the corresponding author on reasonable request.

Authors' contributions: FL, JP, YO, HL, JH, HX, DO, and XW collected and annotated the data, provided clinical knowledge. DY, JW, BW, QS, HF, and YY performed the statical analyses and wrote the manuscript. YX, YL, and XZ supervised and reviewed the study.

Disclosure: **F. Li**, None; **J. Pan**, None; **D. Yang**, (E); **J. Wu**, (E); **Y. Ou**, None; **H. Li**, None; **J. Huang**, None; **H. Xie**, None; **D. Ou**, None; **X. Wu**, None; **B. Wu**, (E); **Q. Sun,** (E); **H. Fang,** (E); **Y. Yang,** (E); **Y. Xu,** (E); **Y. Luo**, None; **X. Zhang**, None

[*] FL, JP, DY, and JW contributed equally to this work.

References

1. Steinmetz Jaimie D., Bourne Rupert RA, Briant Paul Svitil, et al. Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the Right to Sight: an analysis for the Global Burden of Disease Study. *Lancet Global Health*. 2021;9(2):e144–e160.
2. Teo Zhen Ling, Tham Yih-Chung, Yu Marco Chak Yan, et al. Global prevalence of diabetic retinopathy and projection of burden through 2045: systematic review and meta-analysis. *Ophthalmology*. 2021;1285(11):1580–1591.
3. Tham Yih-Chung, Li Xiang, Wong Tien Y., Quigley Harry A., Aung Tin, Cheng Ching-Yu. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology*. 2014;121(11):2081–2090.
4. Wong Wan Ling, Su Xinyi, Li Xiang, et al. Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. *Lancet Global Health*. 2014;2(2):e106–e116.
5. Fricke Timothy R., Jong Monica, Naidoo Kovin S., et al. Global prevalence of visual impairment associated with myopic macular degeneration and temporal trends from 2000 through 2050: systematic review, meta-analysis and modelling. *Br J Ophthalmol*. 2018;102(7):855–862.
6. Ting DS, Pasquale LR, Peng L, et al. Artificial intelligence and deep learning in ophthalmology. *Br J Ophthalmol.* 2019 Feb 1;103(2):167–175.
7. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402, doi:10.1001/jama.2016.17216.
8. De Novo Summary (DEN180001), Available at: https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN180001.pdf. Accessed July 2022.
9. Abràmoff MD, Lou Y, Erginay A, Clarida W, Amelon R, Folk JC, Niemeijer M. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Invest Ophthalmol Vis Sci.* 2016;57(13):5200–5206.
10. He J, Cao T, Xu F, et al. Artificial intelligence-based screening for diabetic retinopathy at community hospital. *Eye*. 2020;34(3):572–576.
11. Sun X, Xu Y, Tan M, Fu H, Zhao W, You T, Liu J. Localizing optic disc and cup for glaucoma

screening via deep object detection networks. *Computational Pathology and Ophthalmic Medical Image Analysis*. Cham: Springer; 2018:236–244.

12. Fu H, Li F, Xu Y, et al.; for iChallenge-GON study group. A retrospective comparison of deep learning to manual annotations for optic disc and optic cup segmentation in fundus photographs. *Transl Vis Sci Technol.* 2020;9(2):33, doi:10.1167/tvst.9.2.33.
13. Li Z, He Y, Keel S, Meng W, Chang RT, He M. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. *Ophthalmology.* 2018;125(8):1199–1206.
14. Burlina PM, Joshi N, Pekala M, Pacheco KD, Freund DE, Bressler NM. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. *JAMA Ophthalmology.* 2017;135(11):1170–1176.
15. Grassmann F, Mengelkamp J, Brandl C, et al.. A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography. *Ophthalmology.* 2018;125(9):1410–1420.
16. Tan Tien-En, Ting Daniel S.W., Liu Yong, et al. Artificial intelligence using a deep learning system with transfer learning to predict refractive error and myopic macular degeneration from color fundus photographs. *Invest Ophthalmol Vis Sci.* 2019;60(9):1478.
17. Du Ran, Xie Shiqi, Fang Yuxin, et al. Deep learning approach for automated detection of myopic maculopathy and pathologic myopia in fundus images. *Ophthalmology Retina*. 2021;5(12):1235–1244.
18. Ting DSW, Cheung CY-L, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*. 2017;318(22):2211, doi:10.1001/jama.2017.18152.
19. Islam MT, Imran SA, Arefeen A, Hasan M, Shahnaz C. Source and camera independent ophthalmic disease recognition from fundus image using neural network. *2019 IEEE International Conference on Signal Processing, Information, Communication & Systems* (SPICSCON). Dahaka, Bangladesh, November 28–30,2019; IEEE:59–63.
20. Lin Duoru, Xiong Jianhao, Liu Congxin, et al. Application of Comprehensive Artificial intelligence Retinal Expert (CARE) system: a national real-world evidence study. *Lancet Digital Health*. 2021;3(8):e486–e495.
21. China Association for Quality Inspection,Zhang X, Xu Y, Yang W Annotation and quality control specifications for fundus color photographs. *Intelligent Medicine*. (2021 Aug 28;1(02):80–7.).
22. Huang Gao, et al. Densely connected convolutional networks. *Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition*. July 26, 2016, Honolulu, Hawaii; 2017.
23. Wu Junde, et al. Gamma challenge: glaucoma grading from multi-modality images. *arXiv preprint arXiv:2202.06511*, 2022.
24. Van Der Heijden A. A., Amber A van der H., Michael D. A., Frank V., Manon V. van H., Albert L., Giel N., et al. Validation of automated screening for referable diabetic retinopathy with the IDx-DR device in the Hoorn Diabetes Care System. *Acta Ophthalmol*. 2018;96(1):63–68.
25. He Jie, Cao T, Xu F, et al. Artificial intelligence-based screening for diabetic retinopathy at community hospital. *Eye*. 2020;34(3):572–576.
26. Yang D., Dalu Y., Yehui Y., Tiantian H., Binghong W., Lei W., Yanwu X., et al. Residual-cyclegan based camera adaptation for robust diabetic retinopathy screening. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Cham: Springer; 2020.
27. Zhang Y., Zhang Y., Ying W., Qingyao W., et al. Collaborative unsupervised domain adaptation for medical image diagnosis. *IEEE Transactions on Image Processing*. 2020;29:7834–7844.