Application of Comprehensive Artificial intelligence Retinal Expert (CARE) system: a national real-world evidence study

Duoru Lin, Jianhao Xiong*, Congxin Liu*, Lanqin Zhao, Zhongwen Li, Shanshan Yu, Xiaohang Wu, Zongyuan Ge, Xinyue Hu, Bin Wang, Meng Fu, Xin Zhao, Xin Wang, Yi Zhu, Chuan Chen, Tao Li, Yonghao Li, Wenbin Wei, Mingwei Zhao, Jianqiao Li, Fan Xu, Lin Ding, Gang Tan, Yi Xiang, Yongcheng Hu, Ping Zhang, Yu Han, Ji-Peng Olivia Li, Lai Wei†, Pengzhi Zhu†, Yizhi Liu†, Weirong Chen†, Daniel S W Ting†, Tien Y Wong†, Yuzhong Chen†, Haotian Lin†*

Summary

Background Medical artificial intelligence (AI) has entered the clinical implementation phase, although real-world performance of deep-learning systems (DLSs) for screening fundus disease remains unsatisfactory. Our study aimed to train a clinically applicable DLS for fundus diseases using data derived from the real world, and externally test the model using fundus photographs collected prospectively from the settings in which the model would most likely be adopted.

Methods In this national real-world evidence study, we trained a DLS, the Comprehensive AI Retinal Expert (CARE) system, to identify the 14 most common retinal abnormalities using 207 228 colour fundus photographs derived from 16 clinical settings with different disease distributions. CARE was internally validated using 21 867 photographs and externally tested using 18 136 photographs prospectively collected from 35 real-world settings across China where CARE might be adopted, including eight tertiary hospitals, six community hospitals, and 21 physical examination centres. The performance of CARE was further compared with that of 16 ophthalmologists and tested using datasets with non-Chinese ethnicities and previously unused camera types. This study was registered with ClinicalTrials.gov, NCT04213430, and is currently closed.

Findings The area under the receiver operating characteristic curve (AUC) in the internal validation set was 0·955 (SD 0·046). AUC values in the external test set were 0·965 (0·035) in tertiary hospitals, 0·983 (0·031) in community hospitals, and 0·953 (0·042) in physical examination centres. The performance of CARE was similar to that of ophthalmologists. Large variations in sensitivity were observed among the ophthalmologists in different regions and with varying experience. The system retained strong identification performance when tested using the non-Chinese dataset (AUC 0·960, 95% CI 0·957–0·964 in referable diabetic retinopathy).

Interpretation Our DLS (CARE) showed satisfactory performance for screening multiple retinal abnormalities in real-world settings using prospectively collected fundus photographs, and so could allow the system to be implemented and adopted for clinical care.

Funding This study was funded by the National Key R&D Programme of China, the Science and Technology Planning Projects of Guangdong Province, the National Natural Science Foundation of China, the Natural Science Foundation of Guangdong Province, and the Fundamental Research Funds for the Central Universities.

Introduction

A retinal examination is important in the detection of both systemic diseases that affect the eye (eg, diabetes and hypertension) and primary ocular diseases (eg, age-related macular degeneration [AMD]).[1] The use of fundus photography based on a teleophthalmology platform is an appealing means to screen and monitor such retinal diseases. The addition of artificial intelligence (AI) to fundus photography provides an opportunity to improve this platform for the detection and monitoring of retinal diseases on a large scale.[2,3] Over the past 5 years, satisfactory performance of AI models for the automated detection of diabetic retinopathy,[4] AMD,[5] and optic-nerve abnormalities[6] from fundus photographs has been reported.

Medical AI technology has moved from the research phase to clinical implementation.[7–9] However, studies that show the performance of image-driven AI deep-learning systems (DLSs) for fundus disease screening in real-world environments are scarce. Real-world evidence has been recommended to be used in clinical evaluation and regulatory decisions about new medical device products.[10] The US Food and Drug Administration authorised the first autonomous AI-based diagnostic system for the detection of diabetic retinopathy after a small-scale clinical trial (900 participants) in primary-care offices in 2018.[11] Another example is a study by Google Health,[7] which used a validated DLS trained on retrospective, well curated retinal images (EyePACS and Messidor)[4] for the

Lancet Digit Health 2021; 3: e486–95

See **Comment** page e463

For the Chinese translation of the abstract see **Online** for appendix 1

*Joint first authors

†Co-senior authors

State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Centre (D Lin PhD, L Zhao MS, Z Li PhD, S Yu MD, X Wu PhD, Prof T Li PhD, Y Li PhD, Prof L Wei PhD, Prof Y Liu PhD, Prof W Chen MD, D S W Ting MD, Prof H Lin PhD) **and Centre for Precision Medicine** (Prof H Lin), **Sun Yat-sen University, Guangzhou, Guangdong, China; Beijing Eaglevision Technology Development, Beijing, China** (J Xiong PhD, C Liu MS, X Hu MS, B Wang MS, M Fu MS, X Zhao MS, X Wang MS, Prof Y Chen PhD); **Department of Electrical and Computer Systems Engineering, Faculty of Engineering, Monash University, Melbourne, VIC, Australia** (Z Ge PhD); **Department of Molecular and Cellular Pharmacology** (Y Zhu PhD) **and Sylvester Comprehensive Cancer Center** (C Chen PhD), **University of Miami Miller School of Medicine, Miami, FL, USA; Beijing Tongren Eye Centre, Beijing Key Laboratory of Intraocular Tumour Diagnosis and Treatment, Beijing Tongren Hospital, Capital Medical University, Beijing, China** (Prof W Wei PhD); **Department of Ophthalmology, Ophthalmology and Optometry Centre, Peking University People's Hospital, Beijing, China** (Prof M Zhao PhD); **Department of Ophthalmology, Qilu Hospital of Shandong University, Jinan, Shandong, China** (Prof J Li PhD); **Department of**

Ophthalmology, People's Hospital of Guangxi Zhuang Autonomous Region, Nanning, Guangxi, China (F Xu PhD); **Department of Ophthalmology, People's Hospital of Xinjiang Uygur Autonomous Region, Urumqi, Shanxi, China** (L Ding MBBS); **Department of Ophthalmology, University of South China, Hengyang, Hunan, China** (G Tan PhD); **Department of Ophthalmology, The Central Hospital of Wuhan, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, China** (Y Xiang PhD); **Bayannur Paralympic Eye Hospital, Bayannur, Inner Mongolia, China** (Y Hu MBBS, P Zhang MBBS); **Department of Ophthalmology, Eye and ENT Hospital, Fudan University, Shanghai, China** (Y Han MBBS); **Moorfields Eye Hospital NHS Foundation Trust, London, UK** (J-P O Li MBBS); **Guangdong Medical Devices Quality Surveillance and Test Institute, Guangzhou, Guangdong, China** (P Zhu PhD); **Singapore Eye Research Institute, Singapore National Eye Centre, Singapore** (D S W Ting MD, Prof T Y Wong MD)

Correspondence to:
Prof Haotian Lin, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, Guangdong 510060, China
gddlht@aliyun.com

or

Prof Yuzhong Chen, Beijing Eaglevision Technology Development, Beijing 100081, China
chenyuzhong@airdoc.com

See Online for appendix 2

Research in context

Evidence before this study

Medical artificial intelligence (AI) has entered the clinical implementation phase, although real-world performance of deep-learning systems (DLSs) for screening fundus disease remains unsatisfactory. A clinically applicable DLS for fundus diseases should be trained by real-world data and externally tested by fundus photographs collected prospectively from the settings in which the model would most likely be adopted. We searched PubMed and Web of Science on June 1, 2020 for articles published between Jan 1, 2010, and May 31, 2020, using the keywords "artificial intelligence", "deep learning", "fundus disease" (or other disease names, including "diabetic retinopathy" and "age-related macular degeneration"), and "real-world", but identified no known studies that tested DLS for fundus diseases using prospectively collected nationwide real-world data. We did not apply any language restrictions. The US Food and Drug Administration authorised the first autonomous AI-based diagnostic system for diabetic-retinopathy detection after a small-scale clinical trial in primary-care offices. Another study was done by Google Health, which used a validated DLS trained on retrospective well curated retinal images for the detection of diabetic retinopathy, was applied in 11 clinics in real-world settings in Thailand. However, the performance of the DLS was well below expectation in a real-world environment, and further model training and testing is required.

Added value of this study

This national real-world evidence study trained a DLS to identify 14 retinal abnormalities using fundus photographs collected from different medical real-world settings and tested the DLS using photographs prospectively collected from settings across China and a series of designed datasets. The model performed well in a real-world environment. This study also showed that an AI solution can be deployed in remote areas with poor network infrastructure and scarce medical resources, while maintaining a high degree of accuracy.

Implications of all the available evidence

Using representative data to train a DLS and testing the model with prospectively-collected real-world data across the country can improve model performance in a real-world environment. This study provides an important reference for the National Medical Products Administration in regulatory decisions about new medical AI-device products.

detection of diabetic retinopathy, and was applied in 11 clinics in a real-world setting in Thailand. The study showed that 21% of fundus photographs in these real-world clinics could not be identified by the system, which affected the overall performance of the DLS model. Li and colleagues[12] also developed and tested a deep-learning algorithm for the detection of glaucomatous optic neuropathy using retinal images downloaded from another public online dataset (LabelMe). Most of these publicly-available photographs are highly selective and have unclear data sources, which are not representative of data in real-world clinical settings. Ideally, the DLS should be tested on fundus photographs prospectively collected from real-world clinical settings, especially from primary-care and community settings with specific environments in which the AI model would most likely be used.[13]

Furthermore, network-connectivity issues are another factor hindering the clinical application of DLS models.[7] Developing a DLS that can detect multiple retinal abnormalities is more valuable for the real-world clinical setting. However, nearly all reported multidisease-identification DLS models were ensembles of multiple binary-classification networks that were separately trained using single disease-labelled photographs.[2,3] Running these complicated DLSs consumes a large amount of the memory of the graphics processing unit (GPU), which requires support from a powerful online GPU server. Developing a simple-architecture DLS with less computational cost to work offline might be more suitable for real-world clinical applications, especially in remote areas with poor networks and scarce medical resources. Finally, the ensembles of DLS models might prevent sharing of disease information and assisted or differential diagnoses among multiple retinal abnormalities, thereby compromising efficiency and hindering further improvement in model performance. Algorithms that are able to detect multiple retinal abnormalities while recognising the correlation between them appear to more closely mimic the thought process of physicians.

Our study aims to address these limitations in the detection of retinal abnormalities using fundus photographs. To the best of our knowledge, this is the first study to use fundus photographs prospectively collected from the real world to test a DLS model for identifying multiple retinal abnormalities in a country with a heterogeneous population.

Methods

Study design

This is a national real-world evidence study involving 51 clinical settings across China (appendix 2 p 2). We developed a Comprehensive AI Retinal Expert (CARE) system to identify 14 common retinal abnormalities using 207 228 colour fundus photographs derived from 16 clinical settings with different disease distributions across China. CARE is a model for a single convolutional neural network (CNN) that is highly efficient with lower computational cost than ensembles of binary DLSs for individual abnormalities. CARE was internally validated using 21 867 photographs and then externally tested using 18 136 photographs prospectively collected from 35 real-world centres across China, where the model

would most likely be adopted, including eight tertiary hospitals, six community hospitals, and 21 physical examination centres. The model performance was compared with that of ophthalmologists from nine provinces of China and compared with ophthalmologists with varying clinical experience. Finally, the validity of CARE was also tested on fundus photographs in non-Chinese ethnicities and previously unused camera types.

40 ophthalmologists licensed in China (each with >5 years of experience) and six retinal experts (each with >10 years of experience) were involved in the annotation process (appendix 2 p 3). Each fundus photograph was randomly assigned to three qualified ophthalmologists for annotation (appendix 2 p 4); if the results were consistent, the annotation was adopted. If the findings were discordant, the three ophthalmologists had to discuss the results to reach a consensus. Expert arbitration was done by three retinal experts if any disagreement occurred in the previous discussion. For each individual, only one fundus photograph per eye was included. In total, 260 830 colour fundus photographs with 45–50° fields of view taken under natural pupil size were used for the training and testing of CARE. All photographs were categorised as normal or labelled with one or more of the following 14 common retinal abnormalities: two ocular manifestations of systemic diseases (referable diabetic retinopathy and referable hypertensive retinopathy) and 12 vision-threatening abnormalities (glaucomatous optic neuropathy, pathological myopia, retinal vein occlusion, retinal detachment, macular hole, macular oedema, central serous chorioretinopathy, epiretinal membranes, retinitis pigmentosa, retinal drusen ≥65 μm, macular neovascularisation, and geographic atrophy). The retinal abnormalities were diagnosed by colour fundus photographs on the basis of a comprehensive consideration of disease characteristics obtained from textbooks, reported literature, and the experience of retinal experts. We summarised the definitions or basis for judgment of the 14 included retinal abnormalities as a reference for the graders (appendix 2 p 25). All retinal abnormalities observed were labelled if several lesions were found in the same photograph.

This study was approved by the institutional review board of the Zhongshan Ophthalmic Centre at Sun Yat-sen University (IRB-ZOC-SYSU). All procedures followed the tenets of the Declaration of Helsinki. All fundus photographs were anonymised and de-identified before the analysis. Informed consent was exempted by the IRB-ZOC-SYSU in the retrospectively collected development and internal validation sets. In the prospectively collected external test set, informed consent was obtained from the patients.

Development and internal validation datasets

CARE was developed using 207 228 fundus photographs retrospectively collected from 16 clinical settings providing different amounts of medical care and disease

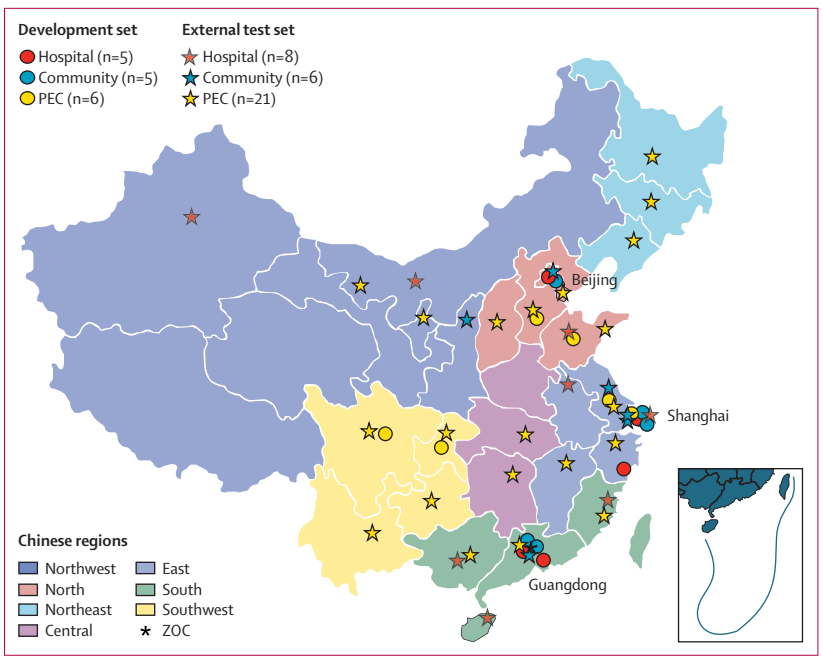

***Figure 1*: Geographical distribution of the clinical settings used in the model training and testing**
The modelling data were predominantly derived from 16 clinical settings in first-tier cities, whereas the fundus photographs included in the external test sets were prospectively collected from 35 settings across China. PEC=physical examination centre. ZOC=Zhongshan Ophthalmic Centre.

distributions across nine provinces or municipalities of China between Jan 4, 2016 and Jun 29, 2018 (appendix 2 pp 16–17). 16 settings with different disease distributions in the development set were mainly located in first-tier cities, including five tertiary hospitals, five community hospitals, and six physical examination centres (figure 1, table 1).

In total, 21 867 fundus photographs derived from the same settings as the development set during a different time period (from July 2, 2018, to Oct 31, 2018) were used to internally validate CARE. The internal validation set was divided into three subgroups according to the data sources as follows: hospital-based datasets with a fundus-disease ratio of 91·21%, community-based datasets with a fundus-disease ratio of 20·96%, and population-based datasets with a fundus-disease ratio of 10·30% (appendix 2 p 18).

External test dataset from 35 centres

The primary aim of CARE was to screen common retinal abnormalities in communities and populations undergoing physical examinations in different regions of China. Therefore, the model was externally tested using 18 136 fundus photographs prospectively collected (from Dec 3, 2018, to Oct 31, 2019) from 35 real-world clinical settings where the model would be used (with no overlap of settings in the development set). The 35 settings are distributed in 28 provinces or municipalities of China (28 [82%] of 34 total provinces and municipalities in China; appendix 2 pp 19–20). The community-based and population-based data were prospectively collected from

	Development	Internal validation set			External test set*			Clinical test set			Total
		Tertiary hospital	Community hospital	Physical examination centre	Tertiary hospital	Community hospital	Physical examination centre	CARE–human competitions	Non-Chinese ethnicity (EyePACS)	Camera-type test	
Photographs	207 228	6735	1614	13 518	3101	7599	7039	358	11 294	1977	260 830†
Labels	253 069	11 107	1732	13 796	4690	7734	7246	366	11 294	1977	313 011
Label of normal fundus	127 508	976	1369	12 375	951	7271	6483	44	3976	1636	162 589
Label of disease	125 561	10 131	363	1421	3645	463	763	322	7318	341	150 328
Type of disease	14	14	9	11	14	9	8	2	1	3	14
Manufacturers of camera	6	6	5	4	6	4	4	4	NA	1	7
Tertiary hospital-based dataset	63 625 (30·7%)	6735 (100%)	0	0	3101 (100%)	0	0	358 (100%)	NA	1977 (100%)	NA
Community hospital-based dataset	53 221 (25·7%)	0	1614 (100%)	0	0	7599 (100%)	0	0	NA	0	NA
Population-based‡ dataset	90 382 (43·6%)	0	0	13 518 (100%)	0	0	7039 (100%)	0	NA	0	NA

Data are n or n (%). The fundus photographs in the development and internal validation datasets were required to be of clinically acceptable quality; more than 80% of the area in the retinal image needed to be easily discriminated, including four main regions (the optic disc, macula, upper-retinal vessel arches, and lower-retinal vessel arches). The images exhibited light leaks covering less than 30% of the area, without spots from lens flares or stains, or severe overexposure. During model training, 211 676 fundus photographs were identified, and 207 228 (97·9%) of 211 676 were included for clinically acceptable quality. During internal validation, 22 482 fundus photographs were identified, and 21 867 (97·3%) of 22 482 were included for clinically acceptable quality (6735 [94·1%] of 7160 in tertiary hospitals; 1614 [98·3%] of 1642 in community hospitals; and 13 518 [98·8%] of 13 680 in physical examination centres). All fundus photographs from the external test set were used in testing CARE. CARE=Comprehensive Artificial intelligence Retinal Expert. NA=not applicable. *A high-performance quality-control model was introduced to assess the quality of fundus photographs before further analysis in the external test set, and 397 (2·2%) of 18 136 photographs were identified as ungradable (184 [5·9%] of 3285 from tertiary hospitals; 130 [1·7%] of 7729 from community hospitals; and 83 [1·2%] of 7122 from physical examination centres). †Including 397 ungradable photographs from the external test set. ‡Fundus photographs were collected from physical examination centres.

Table 1: **Characteristics of the development, internal validation, external, and clinical test sets of CARE**

six community hospitals and 21 physical examination centres across China (figure 1). Given the low prevalence of some retinal abnormalities in the general population, hospital-based data from eight tertiary hospitals (different from those used in the development sets) were also collected to test the model in identification of each of the 14 retinal abnormalities. All fundus photographs from the external test set were used in testing the model (table 1). All fundus photographs were taken by trained non-ophthalmologists. Patients were informed of the primary results immediately after fundus screening by CARE. The final diagnosis reports, confirmed by qualified ophthalmologists, were sent to patients by mobile application and a short message within one day. Blood-sugar concentrations and blood pressure were measured to assist final diagnosis of referable diabetic retinopathy and hypertensive retinopathy by ophthalmologists. Refractive error in dioptres was measured in cases of pathological myopia to ensure diagnostic accuracy.

Algorithm construction of CARE

CARE was trained and tested using InceptionResNetV2 architecture on the TensorFlow platform (version 1.10.1; Google, Mountain View, CA, USA)[14] and the Python scikit learn package 0.22.2. CARE was optimised by the Adam optimiser[15] with an initial learning rate of 0·001. Training and testing were done using GTX 1080Ti GPU ×2 (CUDA version 9.0; Nvidia, Santa Clara, CA, USA) with a batch size of 16. CARE was trained using multidisease-labelled fundus photographs in a single CNN network in which all disease information was shared with each interconnected classifier (figure 2A). This structure was selected as an assembled DLS with independent binary classifiers that do not share disease features (figure 2B). We presented more details regarding the algorithm principle of CARE (appendix 2 p 28).

Clinical tests

We compared the performance of CARE with that of 15 independent binary-classification models trained using 15 single-disease labels (normal fundus and 14 retinal abnormalities). The 15 binary-classification models were trained using the same development set and neural network architecture as CARE. The performance of CARE and the single disease-labelled models was compared using the hospital-based internal test set because of its good coverage of all included retinal abnormalities.

A disease-screening model should perform better than or similarly to physicians before real-world adoption. In this study, the performance of CARE was compared with that of nine ophthalmologists with experience with fundus disease (each 5–10 years of experience in tertiary hospitals) from the following nine provinces or municipalities of China: Beijing in north China; Shandong in north China; Hubei in central China; Hunan in central China; Tibet in northwest China; Xinjiang in northwest China; Guangdong in south China; Guangxi in south China; and Shanghai in East China.

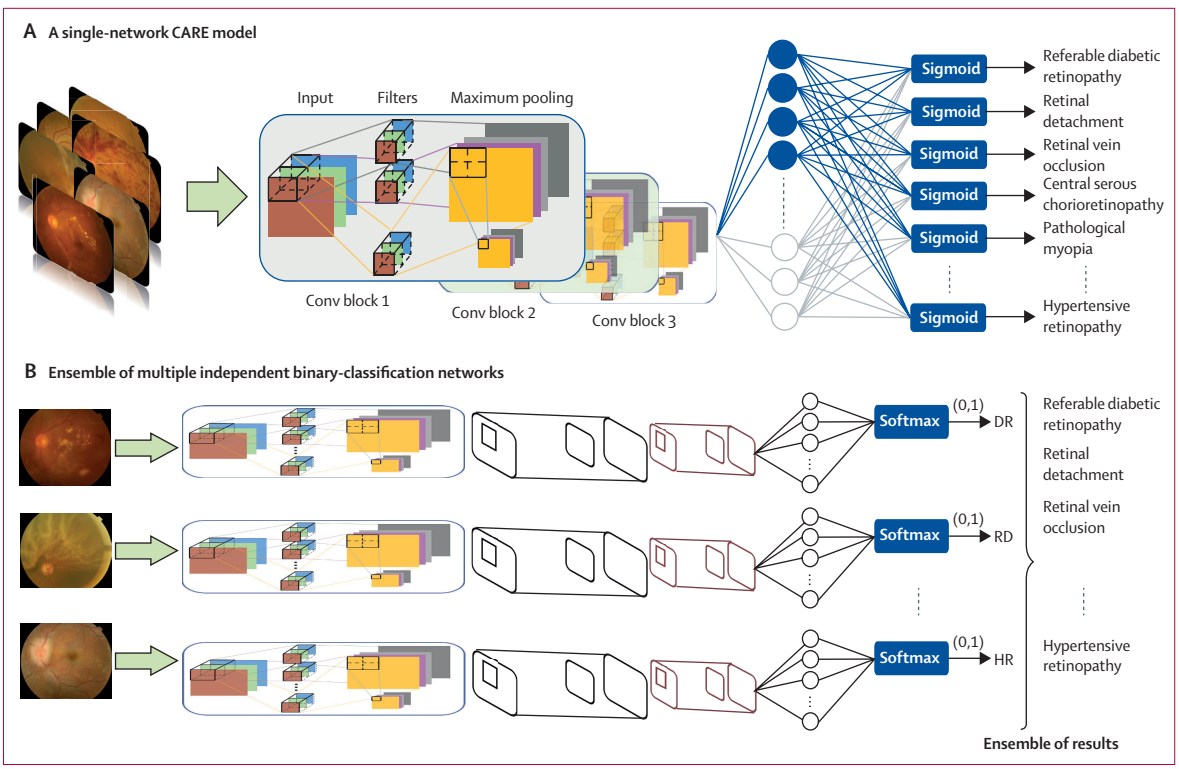

Figure 2: **Algorithm principle comparison between CARE and an assembled deep-learning system**
CARE was trained using multidisease-labelled fundus photographs in a single convolutional neural network in which all disease information was shared with each interconnected classifier (A). This structure was selected as an assembled deep-learning system with independent binary classifiers that do not share disease features (B). CARE=Comprehensive Artificial intelligence Retinal Expert. Conv=convolutional. DR=diabetic retinopathy. HR=hypertensive retinopathy. RD=retinal detachment.

Furthermore, we compared the model performance with that of four groups of Chinese-licensed ophthalmologists with diverse levels of experience as follows: two graduate students with less than 3 years of experience; two ophthalmologists with more than 5 years of experience; two retinal experts with more than 10 years of experience; and one subgroup leader of the retinal-disease group of the Chinese Ophthalmological Society (COS). The dataset used in the comparisons between CARE and the ophthalmologists included 358 additional fundus photographs collected from settings other than those that produced the development and test sets. All of the ophthalmologists involved in the comparisons were different from the 46 annotation doctors or experts.

CARE was tested using datasets with different ethnicities and camera types to validate the model performance. CARE was tested using 11 294 relabelled fundus photographs randomly selected from a public Kaggle dataset (EyePACS LLC, San Jose, CA, USA).[16] The fundus photographs of EyePACS were mostly collected from the Latino population in the USA, and featured different ethnic compositions from the Chinese dataset, including Hispanic patients (nearly 55%), with Black, White, and Asian patients each comprising approximately 5–10% of the population.[4] Because of the small number of disease categories in the EyePACS Kaggle dataset, only the identification of referable diabetic retinopathy (including moderate-to-severe non-proliferative diabetic retinopathy and proliferative diabetic retinopathy) was analysed in this study. As the most common camera types were covered in the development set, CARE was tested using a special dataset of 1977 scanned files of printed fundus photographs from an old-film camera (CR6-45NM, Canon, Tokyo, Japan) that completely differed from those used to capture the development and test sets. Examples of fundus photographs from development, Kaggle (EyePACS), and scanned-file datasets are shown (appendix 2 p 5).

Statistical analysis

All data were stored in the National Engineering Research Centre of Science and Technology Information. The area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and 95% CIs of the AUC of the DLS were calculated to establish and compare model performance. 95% CIs of the AUCs were calculated with 2000 bootstrap samples using the Python scikit learn package (version 0.22.2).[17] This study was registered with ClinicalTrials.gov (NCT04213430).

	Internal validation set			External test set			Threshold*
	Tertiary hospital (n=6735)	Community hospital (n=1614)	Physical examination centre (n=13518)	Tertiary hospital (n=3101)	Community hospital (n=7599)	Physical examination centre (n=7039)	
Referable diabetic retinopathy	0·954 (0·947–0·960)	0·992 (0·983–0·997)	0·852 (0·628–0·999)	0·960 (0·953–0·966)	0·999 (0·998–1·000)	0·918 (0·887–0·944)	0·014
Referable hypertensive retinopathy	0·797 (0·759–0·832)	NA	NA	0·861 (0·788–0·922)	NA	NA	0·019
Glaucomatous optic neuropathy	0·952 (0·945–0·958)	0·968 (0·958–0·977)	0·954 (0·946–0·963)	0·991 (0·989–0·994)	0·993 (0·991–0·996)	0·983 (0·979–0·985)	0·058
Pathological myopia	0·975 (0·970–0·979)	0·993 (0·988–0·996)	0·975 (0·952–0·990)	0·990 (0·986–0·994)	0·995 (0·992–0·997)	0·994 (0·992–0·996)	0·070
Retinal vein occlusion	0·962 (0·959–0·966)	NA	NA	0·948 (0·940–0·956)	NA	NA	0·087
Retinal detachment	0·975 (0·961–0·985)	NA	NA	0·991 (0·970–0·999)	NA	NA	0·025
Macular holes	0·953 (0·932–0·971)	NA	0·999 (0·999–1·000)	0·998 (0·992–1·000)	NA	NA	0·010
Macular oedema	0·975 (0·971–0·978)	0·994 (0·985–0·999)	NA	0·940 (0·933–0·947)	0·999 (0·999–1·000)	NA	0·012
Central serous chorioretinopathy	0·983 (0·976–0·989)	NA	NA	0·974 (0·914–0·999)	NA	NA	0·019
Epimacular membranes	0·951 (0·941–0·960)	0·992 (0·985–0·998)	0·994 (0·990–0·997)	0·934 (0·914–0·952)	0·990 (0·985–0·995)	NA	0·059
Retinitis pigmentosa	0·996 (0·994–0·998)	NA	NA	0·999 (0·999–1·000)	NA	NA	0·018
Retinal drusen	0·916 (0·898–0·932)	0·977 (0·966–0·986)	0·938 (0·867–0·987)	0·948 (0·912–0·975)	0·994 (0·991–0·996)	0·982 (0·971–0·990)	0·006
Macular neovascularisation	0·977 (0·974–0·981)	NA	NA	0·981 (0·973–0·987)	NA	NA	0·113
Geographic atrophy	0·946 (0·910–0·973)	NA	NA	0·999 (0·999–1·000)	NA	NA	0·001
Normal fundus	0·973 (0·969–0·976)	0·903 (0·893–0·914)	0·868 (0·859–0·876)	0·961 (0·956–0·965)	0·908 (0·902–0·915)	0·889 (0·882–0·895)	0·172
Mean AUC (SD)	0·952 (0·045)	0·974 (0·030)	0·940 (0·054)	0·965 (0·035)	0·983 (0·031)	0·953 (0·042)	NA

Data are AUC (95% CI). Model performance in identifying diseases with fewer than five fundus photographs was not analysed. AUC=area under the curve. CARE=Comprehensive Artificial intelligence Retinal Expert. NA=not applicable. *The thresholds are calculated on the basis of tertiary hospital-based data of the internal validation set; the sum of the sensitivity and specificity is maximised to obtain the threshold.

Table 2: **Performance of CARE in internal validation and external tests for retinal abnormalities**

Role of the funding source

The funders had no role in the study design, data collection, data analysis, data interpretation, or writing of this report.

Results

In total, 260 830 fundus photographs were included in the model development and evaluation. The ratio of fundus photographs among the dataset for the model development, internal validation, external test, and clinical test was 80:8:7:5. The disease distributions in the development and test sets are shown (appendix 2 p 18). 397 (2·2%) of 18 136 photographs were classified as ungradable by the algorithm in the external test set (table 1). The information of known manufacturers and types of camera used in this study are presented (appendix 2 p 21).

The performance of CARE using the internal test and external test sets is shown (table 2; appendix 2 pp 6–11). The mean AUC for identifying the 14 retinal abnormalities and normal fundus was 0·955 (SD 0·046) in the internal validation set and 0·968 (0·037) in the external test set. Except for hypertensive retinopathy, CARE exhibited good performance with the included abnormalities, and nearly all AUCs were greater than 0·9. Normal fundus could also be correctly identified by CARE with AUCs ranging from 0·868 (95% CI 0·859–0·876) to 0·973 (0·969–0·976) in datasets with different disease proportions. The anatomical regions that the algorithm might have been using to make its diagnoses were shown by the attention heatmaps (appendix 2 p 12). 2987 (13·66%) of 21 867 fundus photographs in the internal validation set and 863 (4·76%) of 18 136 in the external test set were identified as having multiple abnormalities by CARE; detailed percentage and distribution of multiple diagnoses are shown (appendix 2 p 22).

The comparisons of the model performance between CARE and the single disease-labelled binary models (SBMs) with 14 retinal abnormalities are shown (table 3; appendix 2 p 13). The mean AUC of CARE was higher than that of SBM (0·952 *vs* 0·921) with a narrower SD (0·047 *vs* 0·087). Compared with SBM, CARE was superior at identifying retinal drusen, macular hole, geographic atrophy, and normal fundus. CARE showed marginal superiority in identifying referable hypertensive retinopathy, although neither CARE nor SBM showed satisfactory performance (table 3).

The performance of CARE was similar to that of ophthalmologists in different regions with varying experience (figure 3). Large variations in sensitivity were observed among the ophthalmologists from different regions, ranging from 0·610 to 0·911 in referable diabetic retinopathy and from 0·500 to 0·929 in pathological myopia (appendix 2 p 23). Large variations in sensitivity were also observed among doctors with varying experience, from 0·447 to 0·834 in referable diabetic

	AUC (95% CI)		Sensitivity		Specificity	
	CARE	SBM	CARE	SBM	CARE	SBM
Referable diabetic retinopathy	0·954 (0·948–0·960)	0·973 (0·967–0·978)	0·938	0·951	0·878	0·912
Referable hypertensive retinopathy	0·797 (0·763–0·833)	0·769 (0·745–0·791)	0·600	0·887	0·862	0·562
Glaucomatous optic neuropathy	0·952 (0·945–0·957)	0·970 (0·964–0·976)	0·915	0·960	0·866	0·880
Pathological myopia	0·975 (0·970–0·979)	0·988 (0·984–0·991)	0·898	0·953	0·934	0·953
Retinal vein occlusion	0·962 (0·959–0·966)	0·992 (0·989–0·994)	0·945	0·956	0·905	0·969
Retinal detachment	0·975 (0·962–0·984)	0·917 (0·897–0·936)	0·923	0·936	0·929	0·772
Macular holes	0·953 (0·933–0·970)	0·786 (0·738–0·830)	0·880	0·758	0·869	0·691
Macular oedema	0·975 (0·972–0·978)	0·963 (0·959–0·966)	0·931	0·909	0·924	0·887
Central serous chorioretinopathy	0·983 (0·976–0·989)	0·962 (0·956–0·968)	0·935	0·943	0·933	0·877
Epimacular membranes	0·951 (0·939–0·960)	0·941 (0·933–0·949)	0·882	0·876	0·903	0·877
Retinitis pigmentosa	0·996 (0·994–0·998)	0·991 (0·985–0·997)	0·973	0·959	0·981	0·977
Retinal drusen	0·916 (0·900–0·932)	0·743 (0·733–0·753)	0·860	0·862	0·850	0·556
Macular neovascularisation	0·977 (0·973–0·981)	0·992 (0·990–0·993)	0·922	0·958	0·925	0·953
Geographic atrophy	0·946 (0·913–0·971)	0·869 (0·822–0·911)	0·918	0·754	0·888	0·824
Normal fundus	0·973 (0·969–0·976)	0·962 (0·960–0·965)	0·942	0·943	0·900	0·867
Mean AUC (SD)	0·952 (0·047)	0·921 (0·087)	0·897 (0·087)	0·907 (0·069)	0·903 (0·035)	0·837 (0·136)

The performance of CARE and SBM was compared using the tertiary hospital-based internal validation set (n=6735). AUC=area under the curve. CARE=Comprehensive Artificial intelligence Retinal Expert. SBM=single disease-labelled binary model.

Table 3: **Performance comparisons between CARE and SBM in the identification of 14 retinal abnormalities**

retinopathy and from 0·643 to 0·964 in pathological myopia (appendix 2 p 24). The subgroup leader of retinal disease of COS and the retinal experts exhibited slightly higher performance in disease identification than those with less experience.

CARE showed similar ability in identifying referable diabetic retinopathy (AUC 0·960, 95% CI 0·957–0·964) using the fundus photographs from the Kaggle dataset, which differed from the Chinese dataset in its ethnic composition. Compared with the tertiary hospital-based data of the external test set, CARE exhibited a reduced ability in identifying referable diabetic retinopathy (from AUC 0·960 [0·953–0·966] to AUC 0·882 [0·811–0·945]), geographical atrophy (from AUC 0·999 [0·999–1·000] to AUC 0·899 [0·876–0·920]), and normal fundus (from AUC 0·961 [0·956–0·965] to AUC 0·837 [0·816–0·858]) in the scanned files of fundus photographs from a previously-unused camera type (appendix 2 p 14).

Discussion

A wealth of data is needed for DLS training to discriminate clinically meaningful pathological changes from insignificant features. More still is required for model testing to validate application performance. Using unrepresentative and selective data and other single-centre small-sample databases is not suitable for developing disease-identification DLS models because they limit generalisability and application in real-world clinical environments. In this national real-world evidence study, we trained a DLS (CARE) to identify 14 retinal abnormalities using fundus photographs collected from different medical real-world settings. We tested the model using not only photographs collected prospectively from 35 settings across China where the model would be mostly applied, but also a series of designed datasets with non-Chinese ethnicities and previously unencountered camera types. Furthermore, the requirement of computing power for CARE was relatively low. The computational cost of deploying an online AI service is dependent upon its GPU memory requirements. CARE consumed a maximum memory of 3·6 Gigabytes (GPU RAM) in steady operation. Thus, CARE can be deployed using laptops as a standalone system for large-scale screening even in remote areas with poor networks, which is meaningful for real-world adoption. The module for diabetic retinopathy diagnosis of CARE has been approved by the National Medical Products Administration in China to enter the green channel of innovative medical device applications, and was also part of the first batch of class 3 AI-based devices to be approved for the detection of fundus diseases in China.[18]

Medical AI models are ultimately developed for clinical application and to address unmet clinical needs, especially in community settings. DLSs should be trained by representative data and be clinically tested before their implementation in real-world settings. China is a large and multi-ethnic (56 nationalities) country with 34 provinces and municipalities. CARE development and internal and external tests were based on data from different clinical settings collected from 28 provinces and municipalities, including regions that have the largest number of ethnic groups, which are Yunnan, Inner Mongolia, Ningxia, Xinjiang, and Guangxi. CARE was

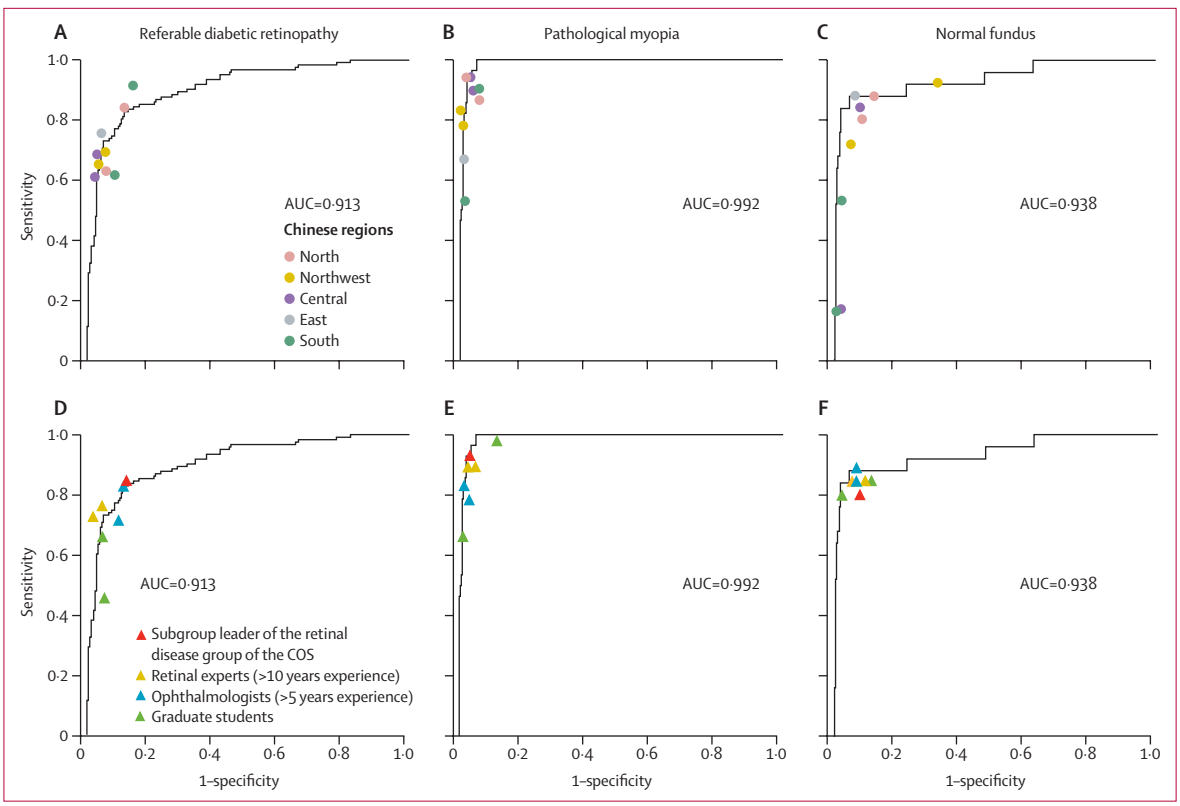

***Figure 3*: Comparisons of model performance with human doctors of various experience in different regions**
The dataset used in the comparisons between CARE and ophthalmologists included 358 additional fundus photographs collected from settings other than those that produced the development and test sets. The performance of CARE was similar to that of ophthalmologists in different regions (A–C) with varying experience (D–F). Large variations in sensitivity were observed among both the ophthalmologists from different regions (ranging from 0·610 to 0·911 in referable diabetic retinopathy and from 0·500 to 0·929 in pathological myopia) and with varying experience (ranging from 0·447 to 0·834 in referable diabetic retinopathy and from 0·643 to 0·964 in pathological myopia). AUC=area under the receiver-operating characteristic curve. CARE=Comprehensive Artificial intelligence Retinal Expert. COS=Chinese Ophthalmological Society.

externally tested using data that was prospectively collected mainly from facilities the model was designed to serve, to approximate the real-world clinical environment of this multi-ethnic country.

We translated the computer logic of diagnosis to imitate the thinking of physicians by training a single-network model that can detect multiple retinal abnormalities using fundus photographs. CARE exhibited a higher AUC with a narrower SD than SBM. CARE enabled connections between coexisting and related pathologies to be made. Patients with pathological myopia[19] and macular oedema[20] are at higher risk for macular hole. Macular hole caused by axial elongation of a pathologically myopic eye is usually complicated by vertical tractional retinal detachment at the posterior pole.[19,21] Retinal drusen, macular neovascularisation, and geographic atrophy are signs of AMD.[22] Retinal drusen appear in any stage of AMD, including macular neovascularisation and geographic atrophy.[23] Macular neovascularisation is less likely to occur with geographic atrophy. Including all these disease labels into one single CNN network might allow CARE to learn the diagnostic logic of retinal abnormalities to achieve further model performance improvements. We found that the abilities of CARE in the identification of retinal drusen and macular hole were improved by nearly 20% compared with those of SBM. Compared with other studies, CARE also exhibited slightly superior or similar AUCs for identifying referable diabetic retinopathy,[2] glaucomatous optic neuropathy,[24] and geographic atrophy.[25]

We further validated the generalisability of CARE by comparing the performance of CARE against that of doctors and fundus photographs derived from patients of non-Chinese ethnicities or taken using previously-unused camera types. Our findings indicate that disease diagnosis by doctors might be more easily affected by their clinical experience and has a risk of misdiagnosis. Because the data used for modelling and testing covered most regions and Chinese ethnicities of China, fundus photographs from Kaggle in the USA with different races from China were selected to test CARE. The ability to identify referable diabetic retinopathy was maintained in the Kaggle dataset, which is similar to those reported in other studies,[4,26] indicating that the performance of CARE does not substantially decrease even when using

fundus photographs of other races. Different camera types have different retinal image characteristics, including tones, exposure times, and pixilation. Given that the most common camera types were included in the model-training dataset, an unseen early type of non-mydriatic retinal film camera was used to test CARE. Scanned files of printed fundus photographs were used to imitate varying photographic quality through differences in colour, exposure, and resolution. Patients usually obtain printed files of their fundus photographs in outpatient departments and can upload electronic copy files taken by telephone cameras to our online diagnosis system (appendix 2 p 15). The performance of CARE decreased to various degrees among referable diabetic retinopathy, glaucomatous optic neuropathy, and normal fundus. The noise of random dots and speckles added into images during scanning could be mistaken by the model for lesions such as microaneurysms and small haemorrhages. Further work is needed to improve the model's performance using images from different camera types.

Our study has limitations. First, cost-effectiveness, patient experience, and clinical practice workflow were not investigated. This study used representative data to train CARE, and the model was tested using data mostly approximating the real-world clinical environment. The model framework was optimised to reduce computational cost, which is central to translating AI models into clinical applications. Second, retinal abnormalities were judged mainly by characteristic features in colour fundus photographs and the clinical experience of retinal experts, some subtle retinal pathologies might have been missed. However, efforts were expanded to test diagnostic accuracy by incorporating data related to blood sugar, blood pressure, and degree of refractive error or related disease history. Third, only 14 representative common retinal abnormalities were included in this preliminary exploration of a multidisease-labelled single-network DLS. Other retinal abnormalities, such as macroaneurysms and retinoblastoma, will be added in our future studies. Furthermore, because of the limitation of traditional fundus photographs with limited visible scope,[27] CARE is not able to identify peripheral retinal pathologies. In addition, only a few diseases were included in the performance comparison between CARE and the ophthalmologists and the tests with non-Chinese ethnicities and unseen camera types. Satisfactory model performance was limited to the tested diseases, and the performance of CARE in the identification of other retinal abnormalities in photographs with previously unused ethnicities and camera types still requires further investigation.

In conclusion, we showed that a DLS (CARE), using a single CNN showed robust performance for the identification of 14 common retinal abnormalities in real-world settings, representing an important development in the journey towards the adoption of AI. CARE was trained using representative fundus photographs and externally tested using data prospectively collected from clinical settings across the country, where the model would be most applied. CARE was also tested through comparisons with single disease-labelled binary-classification models, ophthalmologists, and a series of designed datasets with different ethnicities and camera types. More importantly, the requirement for less computational cost of CARE is important for real-world applications.

Contributors

HL and DL contributed to the concept of the study. HL, DL, JX, and YC critically reviewed the manuscript. HL, DL, JX, CL, LZ, ZL, SY, XWu, and BW designed the study and did the literature search. ZG, XH, MF, XZ, XWa, TL, YoL, WW, MZ, JL, FX, LD, GT, YX, YHu, PZha, YHa, WC, YiL, and PZhu collected the data. DL, JX, LZ, and HL contributed to the design of the statistical analysis plan. DL, HL, LZ, CL, and JX did the data analysis and data interpretation. DL, JX, and HL drafted the manuscript. HL, DL, JX, YZ, CC, J-POL, LW, DSWT, TYW, and YC critically revised the manuscript. HL, DL, and YC provided research funding, coordinated the research, and oversaw the project. All authors had access to all the raw datasets and the corresponding authors (HL and YC) has verified the data and had final decision to submit for publication. All authors reviewed and approved the final manuscript.

Declaration of interests

JX and XZ report a valid patent (application number CN108596895A; for an eye-fundus image-detection method and device based on machine learning as well as system), which is used for the class 3 medical device of the Chinese National Medical Products Administration (registration number 20203210686; fundus-photography auxiliary-diagnosis software for diabetic retinopathy). All other authors declare no competing interests.

Data sharing

Individual participant data will be made available on request, directed to the corresponding author (HL). After approval by the institutional review board of ZOC at Sun Yat-sen University, partial data can be shared through a secure online platform for research purposes. We made use of the open-source machine-learning frameworks TensorFlow and InceptionResNetV2 to do the experiments. Given that many aspects of the experimental system, such as data generation and model training, have a large number of dependencies on internal tooling, infrastructure, and hardware, we are unable to publicly release this code in the current stage. However, all the experiments and implementation details are available in the methods section and appendix 2.

For **TensorFlow** see https://github.com/tensorflow/tensorflow

For **InceptionResNetV2** see https://github.com/tensorflow/models/blob/master/research/slim/nets/inception_resnet_v2.py

Acknowledgments

This study was funded by the National Key R&D Programme of China (2018YFC0116500), the Science and Technology Planning Projects of Guangdong Province (2018B010109008), the National Natural Science Foundation of China (82000946 and 81770967), Natural Science Foundation of Guangdong Province (2021A1515012238), and Fundamental Research Funds for the Central Universities (18ykpy33).

References

1 Liu Y, Wu F, Lu L, Lin D, Zhang K. Videos in clinical medicine. Examination of the Retina. *N Engl J Med* 2015; **373:** e9.

2 Ting DSW, Cheung CY, 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:** 2211–23.

3 Son J, Shin JY, Kim HD, Jung KH, Park KH, Park SJ. Development and validation of deep learning models for screening multiple abnormal findings in retinal fundus images. *Ophthalmology* 2020; **127:** 85–94.

4 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:** 2402–10.

5 Burlina PM, Joshi N, Pacheco KD, Liu TYA, Bressler NM. Assessment of deep generative models for high-resolution synthetic retinal image generation of age-related macular degeneration. *JAMA Ophthalmol* 2019; **137:** 258–64.

6 Milea D, Najjar RP. Artificial intelligence to detect papilledema from ocular fundus photographs. *N Engl J Med* 2020; **382:** 1687–95.

7 Beede E, Baylor E, Hersch F, et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery; Honolulu; April, 2020.

8 Shah P, Kendall F, Khozin S, et al. Artificial intelligence and machine learning in clinical development: a translational perspective. *NPJ Digit Med* 2019; **2:** 69.

9 Chassagnon G, Vakalopoulou M, Paragios N, Revel MP. Artificial intelligence applications for thoracic imaging. *Eur J Radiol* 2020; **123:** 108774.

10 Klonoff DC, Gutierrez A, Fleming A, Kerr D. Real-world evidence should be used in regulatory decisions about new pharmaceutical and medical device products for diabetes. *J Diabetes Sci Technol* 2019; **13:** 995–1000.

11 Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med* 2018; **1:** 39.

12 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:** 1199–206.

13 Lin SY, Mahoney MR, Sinsky CA. Ten ways artificial intelligence will transform primary care. *J Gen Intern Med* 2019; **34:** 1626–30.

14 Abadi M, Agarwal A, Barham P, et al. Tensorflow: large-scale machine learning on heterogeneous distributed systems. *arXiv* 2016; published online March 14. https://arxiv.org/abs/1603.04467 (preprint).

15 Kingma DP, Ba J. Adam: a method for stochastic optimization. *arXiv* 2014; published online Dec 22. https://arxiv.org/abs/1412.6980 (preprint).

16 EyePACS. Why EyePACS. http://www.eyepacs.com/why-eyepacs (accessed Nov 13, 2019).

17 Efron B. Better bootstrap confidence intervals. *J Am Stat Assoc* 1987; **82:** 171–85.

18 National Medical Products Administration. Fundus image-assisted diagnostic software products for diabetic retinopacy appproved. 2020. https://www.nmpa.gov.cn/zhuanti/ypqxgg/gggzjzh/20200810093435157.html?type=pc&m= (accessed Nov 8, 2020).

19 Xin W, Cai X, Xiao Y, et al. Surgical treatment for type II macular hole retinal detachment in pathologic myopia. *Medicine* 2020; **99:** e19531.

20 Pessoa B, Dias DA, Baptista P, Coelho C, Beirao JNM, Meireles A. Vitrectomy outcomes in eyes with tractional diabetic macular edema. *Ophthalmic Res* 2019; **61:** 94–99.

21 Ruiz-Medrano J, Montero JA, Flores-Moreno I, Arias L, Garcia-Layana A, Ruiz-Moreno JM. Myopic maculopathy: current status and proposal for a new classification and grading system (ATN). *Prog Retin Eye Res* 2019; **69:** 80–115.

22 Gheorghe A, Mahdi L, Musat O. Age-related macular degeneration. *Rom J Ophthalmol* 2015; **59:** 74–77.

23 Sakurada Y, Parikh R, Gal-Or O, et al. Cuticular drusen: risk of geographic atrophy and macular neovascularization. *Retina* 2020; **40:** 257–65.

24 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:** 1199–206.

25 Keenan TD, Dharssi S, Peng Y, et al. A deep learning approach for automated detection of geographic atrophy from color fundus photographs. *Ophthalmology* 2019; **126:** 1533–40.

26 Voets M, Møllersen K, Bongo LA. Reproduction study using public data of: development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *PLoS One* 2019; **14:** e0217541.

27 Li Z, Guo C, Nie D, et al. A deep learning system for identifying lattice degeneration and retinal breaks using ultra-widefield fundus images. *Ann Transl Med* 2019; **7:** 618.