nature biomedical engineering

ARTICLES

https://doi.org/10.1038/s41551-021-00745-6

Check for updates

Deep-learning models for the detection and incidence prediction of chronic kidney disease and type 2 diabetes from retinal fundus images

Kang Zhang[1,2,23]✉, Xiaohong Liu[3,23], Jie Xu[4,5,23], Jin Yuan[6,23], Wenjia Cai[6,23], Ting Chen[3]✉, Kai Wang[5], Yuanxu Gao[2], Sheng Nie[7], Xiaodong Xu[5], Xiaoqi Qin[5], Yuandong Su[1], Wenqin Xu[1], Andrea Olvera[1], Kanmin Xue[8], Zhihuan Li[1], Meixia Zhang[1], Xiaoxi Zeng[1,9], Charlotte L. Zhang[10], Oulan Li[10], Edward E. Zhang[10], Jie Zhu[11], Yiming Xu[3], Daniel Kermany[1], Kaixin Zhou[10], Ying Pan[12], Shaoyun Li[13], Iat Fan Lai[14], Ying Chi[15], Changuang Wang[16], Michelle Pei[2], Guangxi Zang[2], Qi Zhang[17], Johnson Lau[18], Dennis Lam[18,19], Xiaoguang Zou[20], Aizezi Wumaier[20], Jianquan Wang[20], Yin Shen[21], Fan Fan Hou[7], Ping Zhang[5], Tao Xu[10]✉, Yong Zhou[22]✉ and Guangyu Wang[5]✉

Regular screening for the early detection of common chronic diseases might benefit from the use of deep-learning approaches, particularly in resource-poor or remote settings. Here we show that deep-learning models can be used to identify chronic kidney disease and type 2 diabetes solely from fundus images or in combination with clinical metadata (age, sex, height, weight, body-mass index and blood pressure) with areas under the receiver operating characteristic curve of 0.85–0.93. The models were trained and validated with a total of 115,344 retinal fundus photographs from 57,672 patients and can also be used to predict estimated glomerular filtration rates and blood-glucose levels, with mean absolute errors of 11.1–13.4 ml min^{-1} per 1.73 m^2 and 0.65–1.1 mmol l^{-1}, and to stratify patients according to disease-progression risk. We evaluated the generalizability of the models for the identification of chronic kidney disease and type 2 diabetes with population-based external validation cohorts and via a prospective study with fundus images captured with smartphones, and assessed the feasibility of predicting disease progression in a longitudinal cohort.

Systemic diseases, including chronic kidney disease (CKD) and diabetes, pose major health care challenges. CKD is a highly prevalent disease and affects approximately 8–16% of the world population[1,2]. CKD is a serious public health problem, as its adverse outcomes are not only limited to end-stage renal failure requiring dialysis or transplantation but also include vascular complications of impaired kidney function[3]. Moreover, cardiovascular events and mortality are strongly associated with CKD in the high-risk diabetic or hypertensive population. Type 2 diabetes mellitus (T2DM) is another major common chronic disease globally, with an estimated prevalence of 9.3% (463 million affected individuals) in 2019. According to the International Diabetes Federation, its prevalence has been increasing steadily in recent years and will reach an estimated 700 million by 2045[4]. According to the US Centers for Disease Control and Prevention, diabetes is one of the leading causes of mortality globally. It is also a leading risk factor for many other common medical problems, including cardiovascular disease, kidney failure and blindness[5–7]. In many of these conditions, early diagnosis and treatment are crucial in reducing the associated comorbidities and mortality. However, early identification and

[1]Center for Clinical Translational Innovations and Biomedical Big Data Center, West China Hospital and Sichuan University, Chengdu, China. [2]Center for Biomedicine and Innovations, Faculty of Medicine, Macau University of Science and Technology and University Hospital, Macau, China. [3]Department of Computer Science and Technology, Tsinghua University, Beijing, China. [4]Beijing Institute of Ophthalmology, Beijing Tongren Eye Center, Beijing Tongren Hospital, Beijing Ophthalmology and Visual Science Key Lab, Beijing, China. [5]State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China. [6]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China. [7]State Key Laboratory of Organ Failure Research, National Clinical Research Center for Kidney Disease and Nanfang Hospital, Southern Medical University, Guangzhou, China. [8]Nuffield Laboratory of Ophthalmology, Department of Clinical Neurosciences, University of Oxford and Oxford University Hospitals NHS Foundation Trust, Oxford, UK. [9]Kidney Research Institute, Nephrology Division, West China Hospital and Sichuan University, Chengdu, China. [10]Bioland Laboratory (Guangzhou Regenerative Medicine and Health Guangdong Laboratory), Guangzhou, China. [11]Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, China. [12]Department of Endocrinology, Kunshan Hospital Affiliated to Jiangsu University, Kunshan, China. [13]The Big Data Research Center, Chongqing Renji affiliated Hospital to the University of Chinese Academy of Sciences, Chongqing, China. [14]Ophthalmic Center, Kiang Wu Hospital, Macau, China. [15]Peking University First Affiliated Hospital, Beijing, China. [16]Peking University Third Affiliated Hospital, Beijing, China. [17]Biotherapy Center, Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China. [18]Department of Applied Biology and Chemical Technology, Hong Kong Polytechnic University, Hong Kong, China. [19]C-MER Dennis Lam and Partners Eye Center, C-MER International Eye Care Group, Hong Kong, China. [20]Ophthalmic Center of the First People's Hospital of Kashi Prefecture, Kashi Prefecture, Xinjiang, China. [21]Medical Research Institute, Wuhan University, Wuhan, China. [22]Clinical Research Institue, Shanghai General Hospital, Shanghai Jiaotong University School of Medicine, Shanghai, China. [23]These authors contributed equally: Kang Zhang, Xiaohong Liu, Jie Xu, Jin Yuan, Wenjia Cai. ✉e-mail: kang.zhang@gmail.com; tingchen@mail.tsinghua.edu.cn; xutao@ibp.ac.cn; yongzhou78214@163.com; guangyu.wang24@gmail.com

diagnosis of CKD and diabetes remain challenging as many patients are asymptomatic or only have non-specific symptoms, with some reports suggesting that up to 5% of the diabetic population remains undiagnosed.

The American Diabetes Association (ADA)[8] recommends annual testing of urinary albumin and estimated glomerular filtration rate (eGFR) for patients who have had diabetes for at least 5 years. A previous study reported that the awareness rate of CKD was less than 20% and the treatment rate was less than 50% in China[9]. Regular screening is crucial for the early detection and diagnosis of CKD and the prevention of its progression. A cost–benefit analysis conducted in China revealed that medical expenses could be considerably reduced by screening for kidney disease in patients newly diagnosed with T2DM[10]. Early detection and intervention are key to the prevention of end-stage renal failure in CKD and sight-threatening complications of diabetic retinopathy.

The retina of the eye is a convenient window into the homeostasis of the body, allowing us to non-invasively observe vascular, neural and connective tissues, both structurally and, in the case of the vasculature, in dynamic action. Systemic disorders may have manifestations in the fundus that allow us to detect, diagnose, stage, monitor and manage the systemic disease. Recent advances using deep-learning classifiers have led to applications of artificial intelligence (AI) in many areas of health care[11–15], including image-based diagnosis[16] and natural language processing[17]. In particular, convolutional neural networks (CNNs) with transfer learning have facilitated efficient and accurate image-based diagnosis well beyond human capabilities[16,18]. There has also been early but promising evidence that identification of systemic conditions and clinical metrics based on fundus photographs is possible[19,20], suggesting that deep-learning algorithms can detect subtle associations that are undetectable to the human observer. Therefore, retinal images, which can be acquired rapidly and non-invasively, may have the potential to provide 'point of care' biomarkers for systemic disease.

Although there is a recent report on CKD detection using a retinal image-based AI system[21], prediction of CKD onset from a normal baseline and diagnosis of early CKD using AI have not yet been described and would have an important impact on disease prevention and favourable outcomes. Similarly, predicting the onset of T2DM would be critical for disease prevention and improving outcomes. Here we explored risk stratification for developing CKD and T2DM using both retinal images and known clinical risk factors. Identifying this asymptomatic pre-morbid population provides the possibility of better channelling health care resources to monitor 'patients at risk' and to modify lifestyle and other risk factors at an early stage.

One of the key practical challenges to the application of AI in health care, particularly in low-resource settings, is the lack of stable computational infrastructure and resources required to run the AI algorithms. Accordingly, the deployment of AI-based technologies through mobile systems has emerged as a growing area of investigation[22]. The latest personal smartphones are equipped with the requisite hardware to run AI algorithms, such as Google Translate, Siri, FaceID and shopping apps with object recognition. With Android's Neural Network API (NNAPI) and the Neural Engine chip in the Apple iPhone X[23], smartphones now deploy a range of machine-learning algorithms including object tracking[24], face detection and recognition[25]. Furthermore, smartphone ownership is widespread worldwide. For instance, in Nigeria, the physician-to-patient ratio is 1:2,660, while smartphone ownership is 1:3.5. An AI-based smartphone diagnostic system is thus an attractive way to broaden health care access by encouraging patients to self-monitor and allowing doctors to diagnose and follow-up with patients remotely.

In this study, we aimed to develop an AI system capable of analysing retinal fundus images to detect CKD and T2DM. We employed deep-learning-based analysis of retinal fundus images for two types of task: a regression task of predicting continuous values (including eGFR) and a binary classification task of making the diagnosis. We also validated the AI algorithm to detect CKD and T2DM in external independent patient populations. Furthermore, we show our AI system can predict disease development and perform risk stratification for CKD and T2DM using retinal images in two longitudinal cohorts (Fig. 1). Using this approach, targeted screening of subgroups of the population could potentially help to deliver risk-reduction interventions to those most likely to develop the diseases. Finally, a smartphone-based system was created to provide a point-of-care system for CKD and T2DM screening in the community.

Results

Definitions of CKD and T2DM. Based on international guidelines and previous studies[1,26], diagnosing an individual with CKD relies primarily on eGFR, an index of kidney function, and renal damage markers (for example, urinary albumin). The presence of CKD is defined by an eGFR $\geq$60 ml min^{-1} per 1.73 m^2 with albuminuria or eGFR <60 ml min^{-1} per 1.73 m^2, confirmed in at least two visits separated by 3 months; healthy controls are defined by an eGFR $\geq$60 ml min^{-1} per 1.73 m^2 without albuminuria. In our study, we utilized images and corresponding eGFR measurements from people who had already been diagnosed with CKD. CKD can be divided into five stages depending on severity[27,28]. For the purpose of this study, we defined three risk stratifications depending on the five stages: early CKD (stages 1 and 2), advanced CKD (stage 3) and severe+ CDK (stages 4 and 5) (Methods).

For the study of T2DM, we utilized images and corresponding clinical data, including laboratory values, from patients who had already been diagnosed with T2DM based on a fasting blood glucose level $\geq$7.0 mmol l^{-1} confirmed in at least two visits, a haemoglobin A1c (HbA1c) value $\geq$6.5% and/or a history of drug treatment for diabetes. The T2DM cohort consisted mainly of individuals with no diabetic retinopathy (NDR) and a small proportion of individuals with diabetic retinopathy (DR) as defined by Early Treatment Diabetic Retinopathy Study (ETDRS) standards[29].

Image datasets and patient characteristics. In the study, a large retinal fundus image dataset encompassing patient cohorts from the China Consortium of Fundus Image Investigation (CC-FII) was constructed, which consisted of cross-sectional datasets and longitudinal datasets. The demographics and clinical information of the cohort participants are summarized in Table 1 and Supplementary Fig. 2.

To develop an AI system for detecting CKD and T2DM, we first used a cross-sectional dataset (CC-FII-C) comprising 86,312 retinal fundus images from 43,156 participants, which is a subset of the CC-FII (Fig. 1a and Table 1). All participants in CC-FII-C were split randomly into mutually exclusive sets for training, tuning and internal testing of the AI algorithm at a 7:1:2 ratio. To evaluate the AI model's generalization, we used two other independent retrospective population-based cohorts for external validation. The first external cohort consisted of non-selected 8,059 individuals who underwent an annual health check from Guangdong Province (Table 1 and Methods).

To further test its generalizability, we conducted a second external test (external test set 2): a population-based prospective study with retinal fundus images captured using a smartphone device. Given the potentially broad appeal of an AI-based medical diagnosis system based on non-invasive retinal imaging, we developed a low-cost hand-held smartphone camera attachment that would enable a health care professional or a patient to capture fundus images for assessment (Supplementary Fig. 12). The images could be uploaded to a Health Insurance Portability and Accountability Act (HIPAA)-compliant cloud service where the AI could autonomously

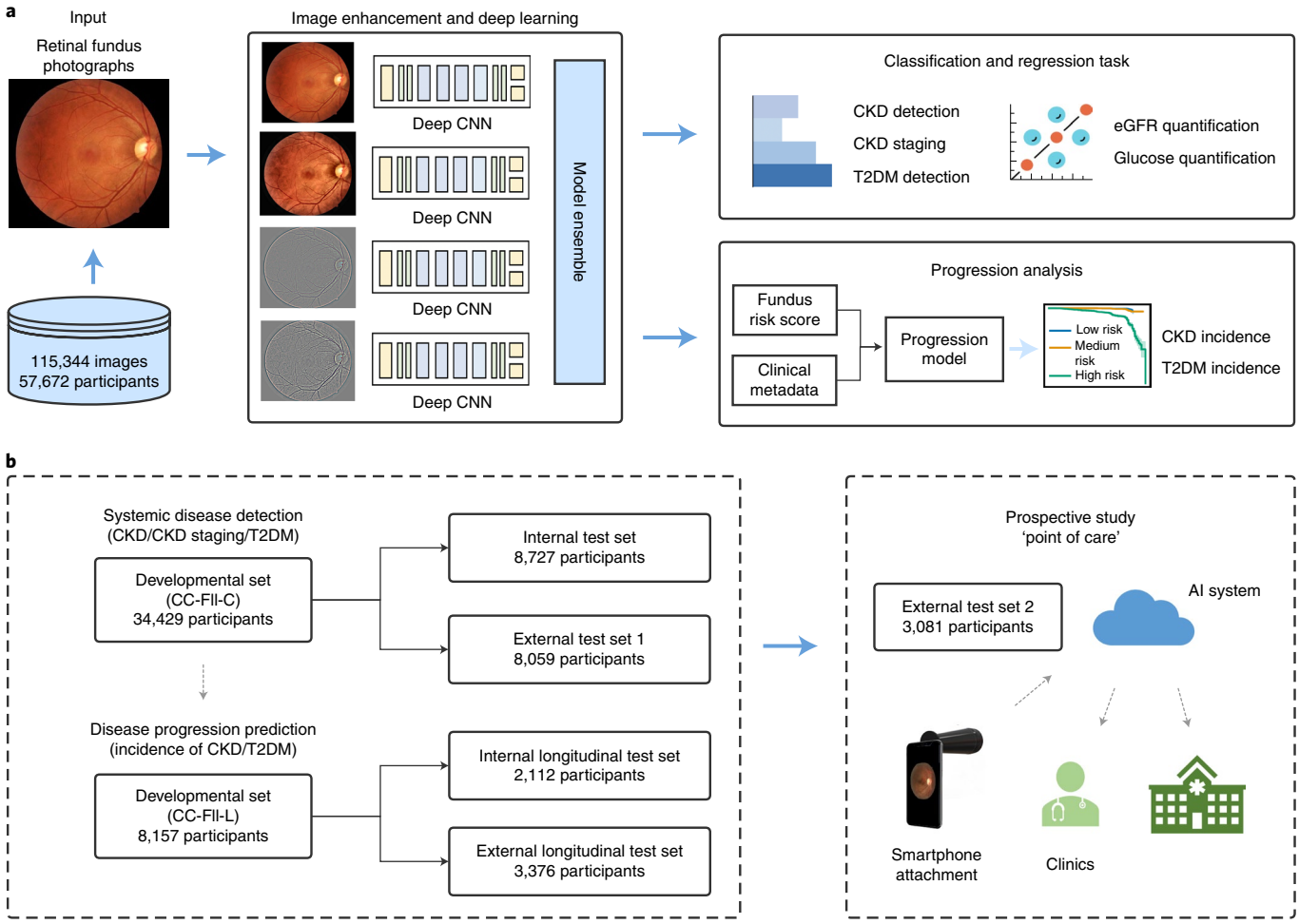

Fig. 1 | AI system for the detection and incidence prediction of systemic diseases using retinal fundus images. a, Illustration of our analysis pipeline. The AI system made two different types of prediction: the assessment of CKD/T2DM and the prediction of disease progression. The AI prediction is generated with an ensemble of model instances. **b**, Evaluation and application of the AI system. Left: training on a developmental dataset for assessment of CKD/CKD staging/T2DM. The model was then tested on independent cohorts to ensure the generalizability. We further show our AI system can predict disease development and perform risk stratification for CKD and T2DM using retinal images in two longitudinal cohorts. Right: a prospective study on a point-of-care setting using a smartphone. Fundus images captured on the phone are transmitted to a cloud hosting the AI model, which produces an instant report transmitted back to the smartphone, an ophthalmologist and a hospital.

grade incoming diagnostic requests. In this prospective study from September to November 2019, 3,081 patients were recruited in the China suboptimal health cohort study (COACS) cohort with 6,162 smartphone-captured fundus images (Supplementary Methods).

To predict the development of CKD and T2DM, we used de-identified fundus images from two longitudinal datasets (Methods). The first dataset (CC-FII-L) for model development is a longitudinal cohort from CC-FII that contained 10,269 individuals from Tangshan City, Hebei Province, China, who underwent routine annual health checks during a six-year follow-up period. The CC-FII-L dataset was randomly split into a developmental dataset and a longitudinal validation set (internal longitudinal test set) at a ratio of 8:2. For external validation, we used the second longitudinal dataset from Beijing, China (external longitudinal test set). This external longitudinal test set contained 3,376 individuals who had annual health check follow-ups for five years. The patient characteristics for each cohort can be found in Supplementary Table 1.

Using the AI system to identify CKD and early CKD. We explored whether an AI algorithm could predict CKD's presence or absence (including early or severe+ stages) in patients based on their fundus images and clinical metadata (for example, age, sex, height, weight and blood pressure; Methods). Based on the definitions of CKD, we trained AI models to perform binary classification tasks. We first developed two models: a baseline random-forest model using clinical metadata and a deep-learning model using fundus images. We hypothesized that a model utilizing both clinical metadata and fundus images might be even more accurate and therefore developed a combined AI model using both input modalities.

Training of an AI model using clinical metadata alone led to an area under the curve (AUC) of 0.861 on the receiver operating characteristic (ROC) curve (95% confidence interval (CI): 0.846–0.877) in the internal test set. In comparison, the AI model trained using fundus images alone produced a superior AUC of 0.918 (95% CI: 0.905–0.933). When trained using combined clinical metadata and fundus images, the combined AI model achieved comparable performance with an AUC of 0.930 (95% CI: 0.921–0.940) (Fig. 2a). We further validated these AI models using another independent external cohort (external test set 1) to demonstrate their generalizability. When tested on external test set 1, the AUCs were 0.842 (95% CI: 0.827–0.856) for the clinical metadata model, 0.885 (95%

Table 1 | Basic characteristics of the patients in the developmental dataset and in the external validation cohorts

Cohorts	Developmental dataset		Internal test set (CC-FII-C)	External test set 1	External test set 2: point of care
	Training set (CC-FII-C)	Tuning set (CC-FII-C)			
Number of images	60,244	8,614	17,454	16,118	6,162
Number of participants	30,122	4,307	8,727	8,059	3,081
Male (%)	15,325 (50.9%)	2,205 (51.2%)	4,412 (50.6%)	4,441 (55.1%)	1,476 (47.9%)
Age (yr)[a]	50.4 ± 14.6	50.6 ± 14.7	50.1 ± 14.6	53.9 ± 13.5	49.0 ± 13.4
BMI (kg m^{-2})[a]	24.7 ± 2.4	24.7 ± 2.4	24.7 ± 2.5	24.8 ± 2.6	24.6 ± 2.7
Hypertension (%)	9,098 (30.2%)	1,308 (30.4%)	2,601 (29.8%)	2,608 (32.4%)	908 (29.5%)
eGFR (ml min^{-1} per 1.73 m^2)[a]	97.1 ± 22.6, n = 19,261	97.0 ± 22.3, n = 2,773	97.6 ± 21.7, n = 5,643	94.7 ± 24.3, n = 4,994	98.3 ± 20.6, n = 3,065
Blood glucose (mmol l^{-1})[a]	6.3 ± 2.6, n = 19,940	6.3 ± 2.7, n = 2,884	6.2 ± 2.7, n = 5,857	7.0 ± 3.1, n = 5,373	6.6 ± 3.0, n = 3,039
CKD (%)	1,906 (17.4%), n = 10,977	251 (16.0%), n = 1,569	484 (15.3%), n = 3,169	676 (14.8%), n = 4,556	228 (18.4%), n = 1,242
Early CKD[b] (%)	648 (34.0%)	82 (32.7%)	159 (32.9%)	240 (35.5%)	71 (31.1%)
Advanced CKD[c] (%)	828 (43.4%)	105 (41.8%)	211 (43.6%)	278 (41.1%)	111 (48.7%)
Severe+ CKD[d] (%)	430 (22.6%)	64 (25.5%)	114 (23.6%)	158 (23.4%)	46 (20.2%)
T2DM (%)	8,791 (29.2%), n = 30,122	1,286 (29.9%), n = 4,307	2,361 (27.1%), n = 8,727	2,823 (35.0%), n = 8,059	1,189 (38.6%), n = 3,081
T2DM-DR (%)	1,414 (16.1%)	228 (17.7%)	392 (16.6%)	425 (15.1%)	141 (11.9%)
T2DM-NDR (%)	7,377 (83.9%)	1,058 (82.3%)	1,969 (83.4%)	2,398 (84.9%)	1,048 (88.1%)

n indicates the number of patients for whom the corresponding measurement/diagnosis was available. [a]Results are presented as mean ± s.d. [b]Early CKD is defined as an eGFR of more than 60 ml min^{-1} per 1.73 m^2 with albuminuria (corresponding to stages 1 and 2). [c]Advanced CKD is defined as an eGFR of 30–59 ml min^{-1} per 1.73 m^2 (corresponding to stage 3). [d]Severe+ CKD is defined as an eGFR of less than 30 ml min^{-1} per 1.73 m^2 (corresponding to stage 4 and above).

CI: 0.873–0.899) for the fundus image model and 0.898 (95% CI: 0.888–0.911) for the combined model (Fig. 2b).

Identification of early CKD is clinically important, as a timely intervention to modify known risk factors (for example, treatment of hypertension, diabetes and urinary tract obstructions) could slow disease progression and reduce long-term health and economic burdens. Early CKD is defined by eGFR ≥60 ml min^{-1} per 1.73 m^2 with albuminuria (a sign of kidney damage), whereas healthy controls have eGFR ≥60 ml min^{-1} per 1.73 m^2 without albuminuria. AI-based prediction of early CKD from healthy controls (in the absence of information about albuminuria) is thus a useful means for early disease detection and prevention. The AI system's performance in predicting early CKD followed a similar trend across the three models (Fig. 2d,e). When evaluated using the internal test set from CC-FII-C, the clinical metadata model achieved an AUC of 0.805 (95% CI: 0.772–0.846), the fundus image model achieved 0.839 (95% CI: 0.805–0.868) and the combined model achieved 0.864 (95% CI: 0.837–0.894) (Fig. 2d). When tested on external test set 1, the AUCs for predicting the presence of early CKD were 0.800 (95% CI: 0.780–0.824) for the clinical metadata model, 0.829 (95% CI: 0.811–0.849) for the fundus image model and 0.848 (95% CI: 0.828–0.869) for the combined model (Fig. 2e).

Despite the AI model having been trained using images captured with standard hospital fundus cameras, the smartphone camera-captured images led to comparably good CKD detection performance based on fundus images alone. For the detection of CKD, the AI delivered AUCs of 0.817 (95% CI: 0.785–0.842) for the metadata-only model, 0.870 (95% CI: 0.847–0.893) for the fundus-only model and 0.897 (95% CI: 0.855–0.902) for the combined model in external test set 2, the point-of-care cohort (Fig. 2c). We further tested the AI performance for the staging of early CKD in the point-of-care study, which achieved AUCs of 0.787 (95% CI: 0.745–0.840), 0.834 (95% CI: 0.797–0.868) and 0.845 (95% CI: 0.812–0.892) for the metadata-only, fundus-only and combined models, respectively (Fig. 2f). Together, these findings demonstrated not only the validity of the AI model, but also the potential real-life feasibility and utility of an AI-based retinal diagnosis administered via personal mobile devices. This could assist medical professionals in screening and monitoring systemic microvascular diseases.

eGFR prediction and CKD staging using retinal fundus images. eGFR is normally calculated based on measurement of serum creatinine level (Methods). We tested whether a patient's eGFR from measured serum creatinine could be predicted by an AI model using their fundus images alone. A linear correlation between the AI-predicted GFR and measured eGFR showed strong associations with a coefficient of determination (R^2) of 0.507, Pearson's correlation coefficient (PCC) of 0.716 and mean absolute error (MAE) of 11.1 for internal test set 1 (from CC-FII) (Fig. 3a). When evaluated on the external validation cohorts, the AI model achieved R^2 of 0.481, PCC of 0.700 and MAE of 12.9 for external test set 1 and R^2 of 0.327, PCC of 0.577 and MAE of 11.8 for external test set 2 (Fig. 3b,c). These results suggest that the AI model was able to extract information predictive of eGFR, the key index of renal function, embedded subtly within fundus images.

The level of agreement between the algorithm-predicted GFR and measured eGFR was assessed using a Bland–Altman plot. The AI model demonstrated good performance in eGFR prediction in the internal test set, achieving an intraclass correlation coefficient (ICC) of 0.65 (95% CI: 0.63–0.66) (Fig. 3d). The model performed similarly well on the first external test set with an ICC of 0.62 (95% CI: 0.60–0.64) (Fig. 3e). When tested on the second external test cohort using smartphone-captured images, the AI model achieved non-inferior performance with an ICC of 0.53 (95% CI: 0.50–0.55) (Fig. 3f). Bland–Altman plots showed a negative proportional bias (slope of linear fit); that is, the AI-based model underestimated the eGFR level to a greater extent at high levels of eGFR than at low levels. The models were trained by minimizing the mean-square error (MSE) loss, resulting in a smaller output variance than the variance of the ground-truthed data. This proportional bias (slope

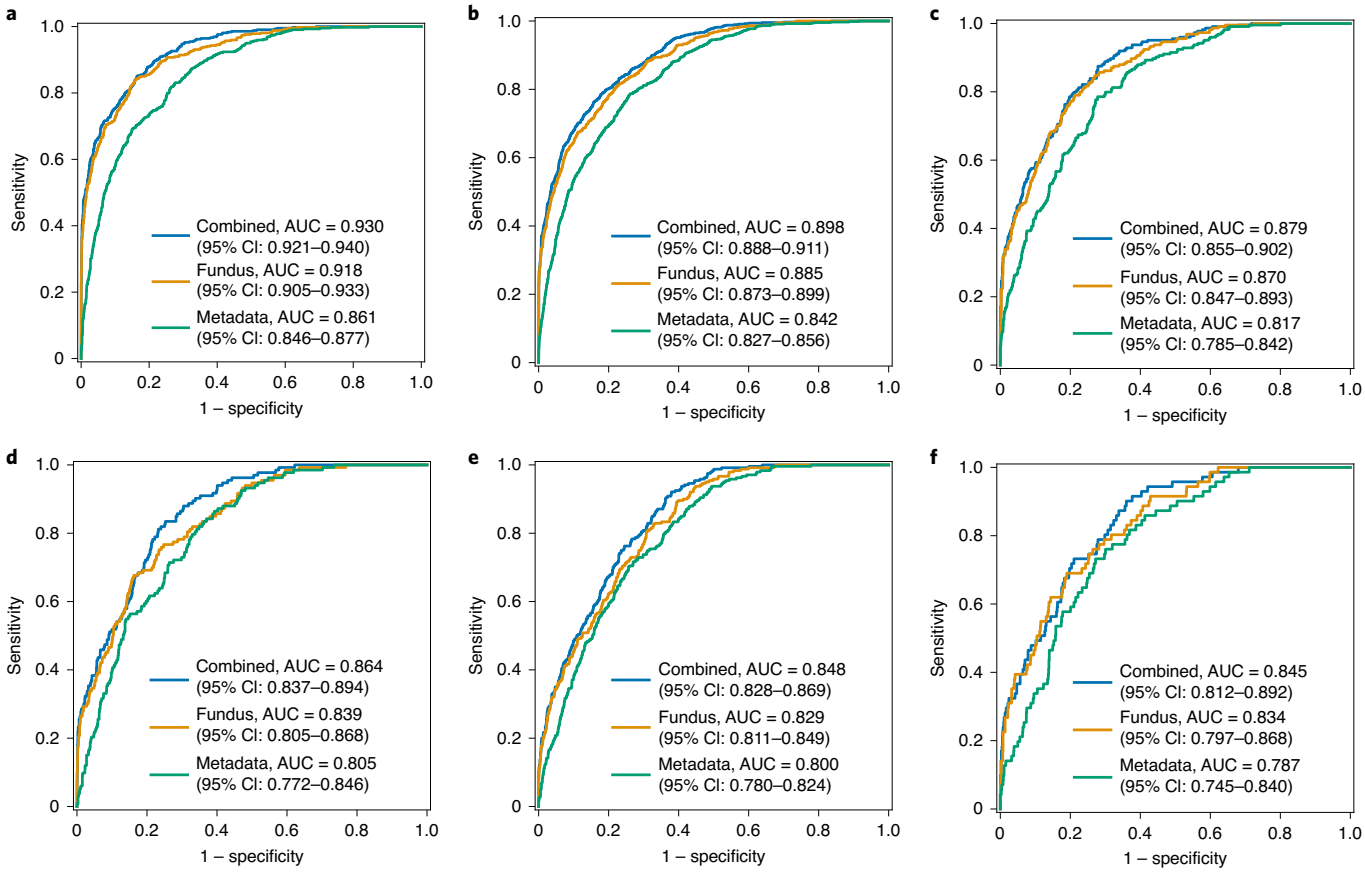

Fig. 2 | Performance of the AI models in the identification of CKD and early CKD. ROC curves represent the metadata-only model, the fundus-only model and the combined model. **a**–**c**, ROC curves showing performance of CKD detection on: internal test set (case:control = 484:2,685) (**a**); external test set 1 (case:control = 676:3,880) (**b**); and external test set 2 (case:control = 228:1,014), a prospective point-of-care pilot study using retinal fundus images from a smartphone (**c**). **d**–**f**, ROC curves showing performance of early CKD detection on: internal test set (case:control = 159:2,685) (**d**); external test set 1 (case:control = 240:3,880) (**e**); and external test set 2 (case:control = 71:1,014), a prospective point-of-care pilot study using retinal fundus images from a smartphone (**f**).

of linear fit) was reduced after the model outputs were calibrated to have the same variance as that of the ground-truthed measurements (Supplementary Fig. 3a–c).

Early detection and treatment of advanced to severe CKD could slow or prevent progression to end-stage renal failure and reduce mortality[2]. With the purpose of fundus imaging-based population screening of CKD in mind, we examined the AI model's performance in predicting the stage of CKD. A regression model was first trained, generating a binary output in terms of presence or absence of severe+ CKD by applying a threshold after predicting the eGFR. Alternatively, a classification model was also trained to perform the binary task of directly differentiating severe+ CKD from the other stages of CKD (early and advanced CKD), which achieved good performance with an AUC of 0.853 (95% CI: 0.799–0.891) on the internal validation dataset A (Supplementary Fig. 3d). Evaluated on the internal test set, the regression model showed comparable performance to the classification model in the detection of severe+ CKD with an AUC of 0.825 (95% CI: 0.776–0.867) (Supplementary Fig. 3d). Similarly, when evaluated on the first external test set, the AUC for severe+ CKD identification was 0.842 (95% CI: 0.803–0.892) for the direct classification model and 0.837 (95% CI: 0.788–0.877) for the indirect regression model (Supplementary Fig. 3e).

Prediction of the development of CKD using longitudinal cohorts. Accurate prediction of CKD development has important implications for delivering targeted screening programmes or risk-modifying interventions to those who would derive the most benefit from early disease detection. We hypothesized that an AI algorithm could predict not only the current CKD status of a patient based on their fundus images but also their future risk of CKD onset or progression. In this study, we implemented a deep-learning-based AI model using baseline fundus images and clinical metadata to predict risk of progression to CKD/advanced CKD in longitudinal cohorts (Supplementary Table 1).

The performance of progression prediction models of CKD/advanced CKD using the Cox proportional hazards (CPH) model has been summarized in Supplementary Table 5. The metadata-based model achieved a *C*-index of 0.756 (95% CI: 0.699–0.810) on the internal test set and a *C*-index of 0.651 (95% CI: 0.569–0.730) on the external test set. When the risk scores extracted from fundus images were combined with clinical metadata, the model performance improved to a *C*-index of 0.845 (95% CI: 0.789–0.910) on the internal test set and of 0.719 (95% CI: 0.627–0.807) on the external test set. The combined progression prediction model had a statistically significant improvement compared with clinical metadata only-based prediction, as shown in Supplementary Table 5 (permutation test). In addition, we conducted analysis on the univariable and multivariable hazard ratio (HR) comparing the fundus-based models to known risk factors (Supplementary Table 3). We used HRs to examine and qualify the influence of specific factors on the rate of occurrence of a particular event rate (for example, onset of disease) at a particular point in time. As shown in

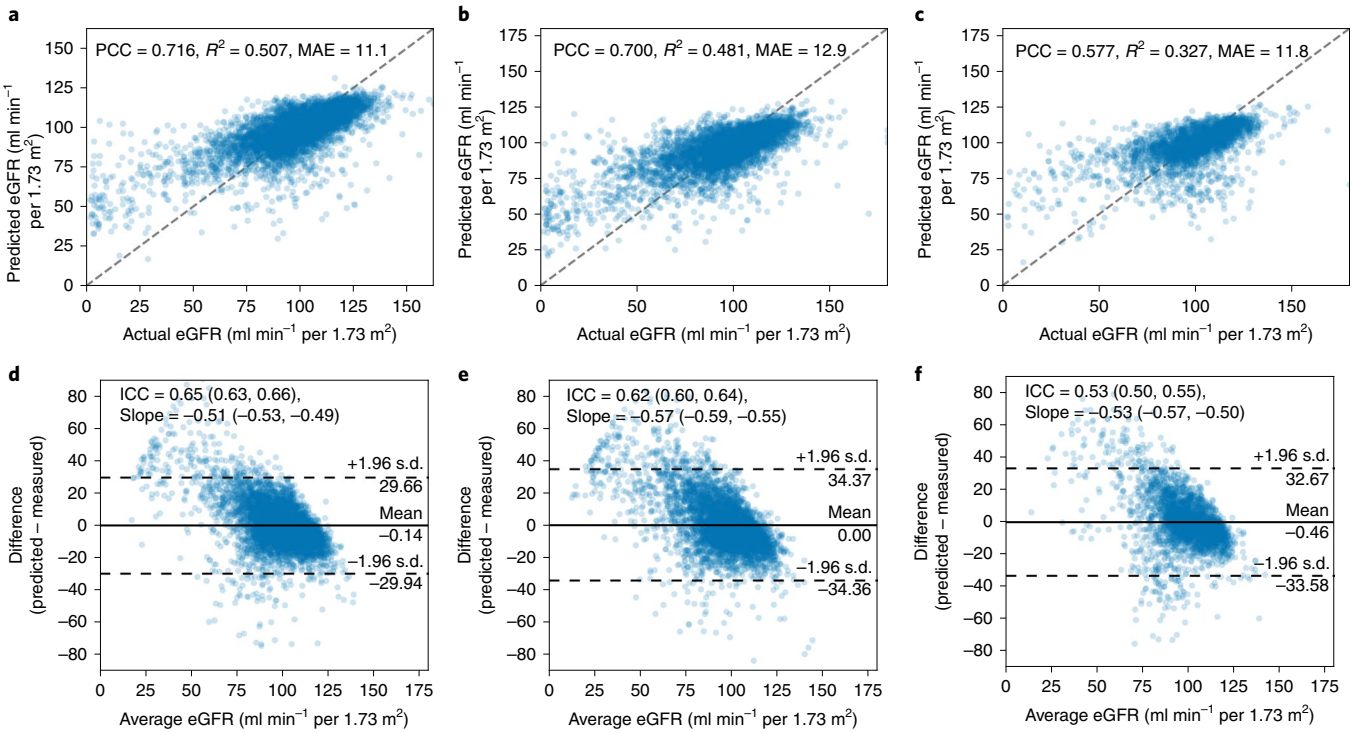

Fig. 3 | Model performance in assessing eGFR from retinal fundus images. a–c, Correlation analysis of the predicted GFR versus actual eGFR generated using the regression model. The performance of the AI system on the internal test set (**a**), external test set 1 (**b**) and external test set 2: point-of-care study (**c**). **d–f**, Bland–Altman plots for the agreement between the predicted GFR and eGFR. The *x* axis represents the mean of predicted GFR and eGFR, and the *y* axis represents the difference between the two measurements. The performance of the AI system on the internal test set (**d**), external test set 1 (**e**) and external test 2, a prospective point-of-care pilot study (**f**).

Supplementary Table 3, the adjusted HR of the fundus predictor is significant ($P < 0.001$).

Using the six-year longitudinal data on CKD staging from our developmental patient cohorts (CC-FII-L), we further used the Kaplan–Meier method to stratify healthy individuals into three risk groups (low, medium or high risk) for developing CKD or advanced to severe CKD (advanced+CKD). The incidence of CKD or advanced+ CKD (per 1,000 person-years) stratified by risk groups of the AI model is shown in Table 2 and Supplementary Table 4. For the Kaplan–Meier curves and log-rank tests, thresholds for the high-risk and low-risk groups were based on the upper and lower quartiles of the predicted risk scores from the combined models in the developmental set. Significant separations of the low-, medium- and high-risk groups were also achieved in the internal test set ($P < 0.001$, Fig. 4a,c and Supplementary Table 7). To assess the generalizability of the AI model, the same tests were performed on an external longitudinal test set of 3,376 patients with five years of follow-up data. Once again, the AI model discriminated the low- versus medium- and high-risk groups for developing CKD or advanced+ CKD with high degrees of separation ($P < 0.001$ for both, Fig. 4b,d and Supplementary Table 7).

We further compared the combined model and metadata model using the same participants in the internal/external test sets, as well as the longitudinal test sets. The Kaplan–Meier curves for the risk stratification from the metadata-only model are illustrated in Supplementary Fig. 7. Supplementary Fig. 8 presents the cumulative hazards of three stratified risk groups in progressing to CKD/advanced CKD outcome at every time point. As shown, the combined model was discriminative in stratifying patients into low-, medium- and high-risk subgroups on both the internal longitudinal test set ($P < 0.001$) and the external longitudinal test set ($P < 0.001$).

In addition, we assessed the prognostic accuracy of the AI system using time-dependent ROC analysis. The AUC for predicting the development of CKD was 0.844 (95% CI: 0.787–0.888) on the internal test set. We further used the AUC at four years in the external longitudinal set to measure prognostic accuracy, which was 0.771 (95% CI: 0.677–0.840) for predicting the onset of CKD (Supplementary Fig. 9a,b).

Using the AI system to identify T2DM. To extend the scope of our AI system in the diagnosis and prediction of systemic microvascular diseases based on fundus images, we applied the same developmental framework to the detection of T2DM. The developmental dataset was divided into training, tuning and internal test sets (at a ratio of 7:1:2) to assess the models' performance (Supplementary Table 1).

We first evaluated our models' performance in the detection of T2DM patients from healthy controls. The AI system achieved an AUC of 0.828 (95% CI: 0.814–0.841) for the metadata-only model, an AUC of 0.923 (95% CI: 0.913–0.932) for the fundus image-only model and an AUC of 0.929 (95% CI: 0.920–0.937) for the combined model on the internal test set (Fig. 5a). When evaluated on the first external test set, the model performed well, with AUCs of 0.796 (95% CI: 0.779–0.814) for the metadata-only model, 0.854 (95% CI: 0.839–0.871) for the fundus-only model and 0.871 (95% CI: 0.856–0.885) for the combined model (Fig. 5b). When evaluated on the second external test set using smartphone camera-captured images, the AI model achieved comparably good T2DM detection performance with AUCs of 0.762 (95% CI: 0.732–0.786) for the metadata-only model, 0.820 (95% CI: 0.788–0.853) for the fundus-only model and 0.845 (95% CI: 0.822–0.869) for the combined model (Fig. 5c).

Table 2 | Predicted incidence rates of CKD/T2DM (per 1,000 person-years) for the internal longitudinal test set and for the external longitudinal test set, stratified by risk level

Subset	Number of participants	Number of events	Incidence rate (95% CI)	Univariate analysis		Multivariate analysis	
				HR (95% CI)	*P* value	HR (95% CI)	*P* value
Prognostic analysis: CKD (internal longitudinal test set)							
Low risk	460	5	2.4 (0.8, 5.7)	0.65 (0.33, 1.28)	0.212	1.08 (0.50, 2.35)	0.847
Medium risk	815	13	3.6 (1.9, 6.2)	Reference	NA	Reference	NA
High risk	410	62	39.0 (29.9, 50.0)	6.69 (4.17, 10.72)	<0.001	2.70 (1.44, 5.10)	0.002
Overall[a]	1,685	80	11.1 (8.8, 13.8)	3.44 (2.70, 4.37)	<0.001	1.87 (1.32, 2.65)	<0.001
Prognostic analysis: CKD (external longitudinal test set)							
Low risk	397	10	6.5 (3.1, 12.0)	0.88 (0.49, 1.61)	0.689	1.12 (0.56, 2.23)	0.750
Medium risk	1,145	22	5.3 (3.3, 8.0)	Reference	NA	Reference	NA
High risk	342	34	29.2 (20.2, 40.8)	4.35 (2.70, 7.02)	<0.001	2.51 (1.28, 4.93)	0.008
Overall[a]	1,884	66	9.6 (7.4, 12.2)	3.73 (2.42, 5.73)	<0.001	2.21 (1.29, 3.79)	0.004
Prognostic analysis: T2DM (internal longitudinal test set)							
Low risk	476	1	0.5 (0.0, 2.8)	0.26 (0.13, 0.51)	<0.001	0.40 (0.19, 0.86)	0.019
Medium risk	839	34	10.5 (7.3, 14.7)	Reference	NA	Reference	NA
High risk	463	54	29.8 (22.4, 38.9)	2.93 (1.97, 4.37)	<0.001	1.66 (0.93, 2.95)	0.084
Overall[a]	1,778	89	12.6 (10.2, 15.6)	4.08 (2.67, 6.23)	<0.001	1.46 (0.77, 2.80)	0.249
Prognostic analysis: T2DM (external longitudinal test set)							
Low risk	430	1	0.6 (0.0, 3.4)	0.27 (0.14, 0.51)	<0.001	0.35 (0.17, 0.69)	0.003
Medium risk	1,600	57	10.4 (7.9, 13.4)	Reference	NA	Reference	NA
High risk	1,114	133	36.0 (30.2, 42.7)	3.22 (2.43, 4.26)	<0.001	2.01 (1.37, 2.96)	<0.001
Overall[a]	3,144	191	17.7 (15.2, 20.3)	3.31 (2.30, 4.77)	<0.001	1.90 (1.17, 3.07)	0.009

[a]A continuous variable was used (predicted *z*-score). *P* values were computed using one-sided likelihood ratio tests for the hazard ratios. NA, data not available.

As a T2DM individual with DR can be readily detected using retinal images, we tested our AI performance independent of DR. Retinal fundus images of T2DM patients were divided into two subsets: (1) T2DM patients with DR and (2) T2DM patients with NDR. Comparable performance of the AI model in detecting T2DM with DR or with NDR suggests that its accuracy is not heavily dependent on the presence of ETDRS-defined DR. These results indicate that the AI model could detect T2DM based on fundus images before any apparent clinical manifestations of DR (Supplementary Fig. 5).

We further tested the ability of our models to predict the mean blood glucose level from fundus images alone. Interestingly, the AI model achieved a relatively strong performance on the internal test set, external test set 1 and external test set 2, suggesting that blood glucose levels could be predicted and quantified from fundus images alone (Fig. 5d–f). Bland–Altman analysis and relating calibration were also performed to investigate the agreement between two approaches of the blood glucose measurement (Supplementary Fig. 4).

Prediction of the development of T2DM using longitudinal cohorts. We next investigated the predictive performance of the AI system for the development of T2DM in healthy individuals over a five-year period. The healthy individuals within the developmental cohort were stratified into low-, medium- or high-risk groups for developing T2DM as defined by the upper and lower quartiles of our AI prediction. The incidence of the T2DM (per 1,000 person-years) stratified by the three risk groups of the AI model is shown in Table 2. As shown in the Kaplan–Meier survival curve, a clean stratification of T2DM development rates between the low-risk and high-risk groups was observed in both the internal longitudinal test set ($P<0.001$) and the external longitudinal test set ($P<0.001$) (Fig. 5g, h and Supplementary Table 7).

Subsequently, we performed progression analysis of the T2DM using CPH models. The performance of the metadata-based model and combined model is summarized in Supplementary Table 6. The metadata-based model achieved a *C*-index of 0.774 (95% CI: 0.732–0.819) on the internal test set and a *C*-index of 0.746 (95% CI: 0.706–0.775) on the external test set. When the risk scores extracted from fundus images were combined with clinical metadata, the model performance improved to a *C*-index of 0.781 (95% CI: 0.743–0.819) on the internal test set and 0.765 (95% CI: 0.723–0.799) on the external test set. We further performed univariable and multivariable survival analyses, including the basic prognostic factors, age, sex and height, in addition to the scores generated from the T2DM detection (Supplementary Table 3).

Among a total of 3,144 participants with verified disease development outcomes, 191 (6.1%) developed T2DM within a five-year follow-up period. The classical approach of ROC curve analysis considers an individual's event (disease) status as fixed over time; however, in practice, the disease status may change over time. In this study, we used time-dependent ROC curve analysis to assess the predictive ability of fundus image features. The AUC for predicting the onset of T2DM was 0.839 (95% CI: 0.799–0.886) on the internal longitudinal test set and 0.824 (95% CI: 0.793–0.858) on the external longitudinal test set (Supplementary Fig. 9c, d).

Visualization of evidence for CKD and T2DM prediction. Finally, to improve the interpretability of the AI model and shed light on its diagnostic mechanism, integrated gradients was used to generate saliency maps to highlight areas of the images that were important in determining the AI model's predictions. Several representative examples of original fundus images and their corresponding saliency maps are presented in Fig. 6. While microvascular pathologic changes are known to exist in CKD and T2DM, they may not

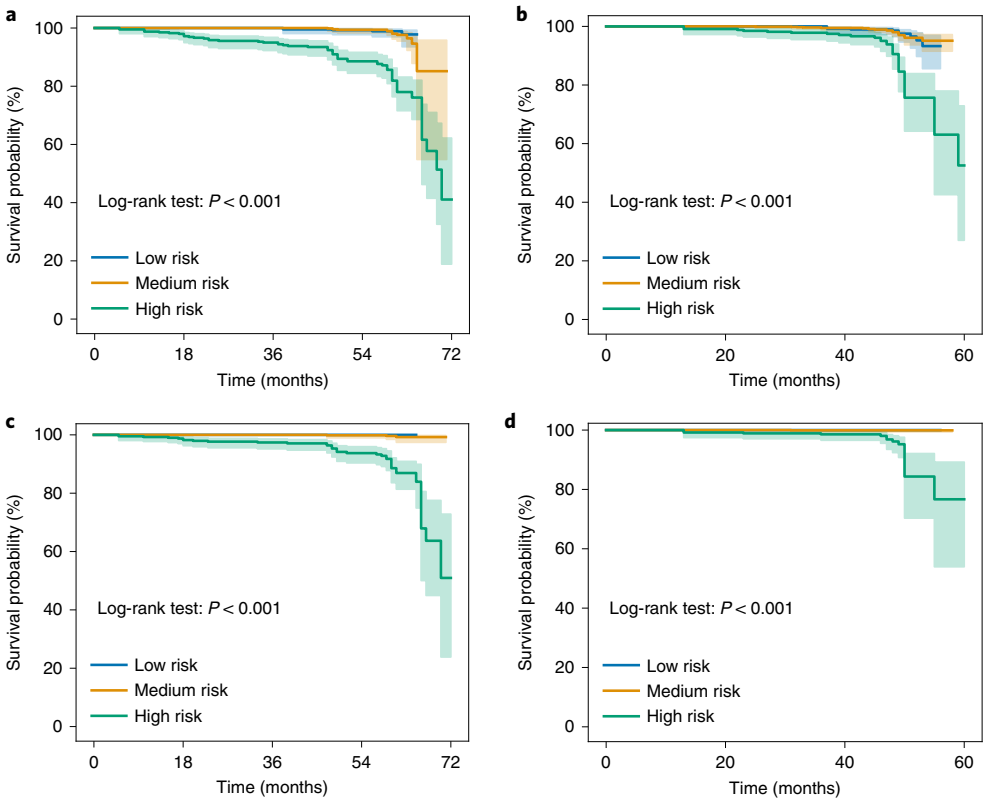

Fig. 4 | Kaplan–Meier plots for the prediction of CKD and advanced+ CKD development. The *y* axis is the survival probability, measuring the probability of not progressing to a disease outcome. The *x* axis is the time in months. Survival curves in different colours represent the high-risk, medium-risk and low-risk subgroups stratified by the upper and lower quartiles in the tuning dataset. Shaded areas are 95% CIs. **a**,**b**, The incidence of CKD in the internal longitudinal test set (**a**) and the external longitudinal test set (**b**). **c**,**d**, The incidence of advanced+ CKD (corresponding to stage 3 or more severe) in the internal longitudinal test set (**c**) and the external longitudinal test set (**d**). *P* value is computed using a one-sided log-rank test between all groups.

be observable by ophthalmologists in retinal fundus photographs. The knowledge derived from saliency maps suggests that the AI model focuses on areas around the optic disc, central macula, retinal vessels and other specific areas of abnormalities (for example, in cases of DR). These areas are commonly used by ophthalmologists to diagnose retinal diseases, suggesting that the AI model is learning clinically relevant features. Interestingly, the saliency maps for CKD and T2DM show somewhat distinct patterns of emphasis. In CKD, the predictive value locations appear to be the optic nerve, vessel branch points and arterial–venous junctions, suggesting that vascular health may play a role (Fig. 6a–c). In T2DM, the highlighted points of interest appear more scattered over the whole image and, in some instances, appear to correspond to the locations of DR features such as vascular tortuosity, venous dilatation, retinal haemorrhage and cotton wool spots (Fig. 6d,e).

Discussion

The common systemic microvascular diseases, CKD and T2DM, are major public health burdens, and early detection is critical for successful clinical management and favourable outcomes. We hypothesized that an AI model could diagnose CKD and T2DM by isolating and identifying subclinical changes that are undetectable by human observers. In this study, we developed an AI to detect CKD and T2DM with high degrees of accuracy and predict future disease development based on fundus images alone and before any clinical manifestation. These AI-derived diagnostic capabilities were applicable not only to clinical-grade retinal images obtained using professional cameras but also to fundus images captured using smartphones. These findings could potentially provide a non-invasive, high-throughput and low-cost screening tool for early detection of CKD and T2DM during health screening. Additionally, the added value of fundus images in disease prognostication was further validated. The results raise the possibility of AI-based detection of other systemic diseases with retinal manifestations beyond clinicians' observational power.

Although the number, distribution and quality of the studies examining CKD prevalence and incidence have increased over the past decade, CKD surveillance capacity remains far less developed, with only 12.1% of patients identified as having CKD by primary care practitioners. Awareness of CKD remains low at 10% in US adults in part because CKD is usually a silent condition until its late stages[30]. In this regard, a recent report on CKD diagnosis using fundus images raises possibilities for early disease detection[21]. Earlier detection and prediction of disease development remains challenging but has great potential to help target interventions to individuals at high risk, thus reducing personal and economic burdens while improving quality of life. Given the prevalence of CKD, as well as an enormous screening demand, our AI system, which is capable of both diagnosing the disease and predicting the risk of onset, could improve CKD surveillance and identification of patients for early intervention. In this regard, the AI system should be viewed as a welcome addition to assist physicians and better direct health care resources.

In addition, diabetes is a major global risk factor for developing high-risk CKD[1]. Good control of blood glucose levels reduces the risk of CKD and improves outcomes in patients with CKD[31]. Our retinal fundus image-based AI system showed its value for T2DM screening in general clinics and communities. As the majority of

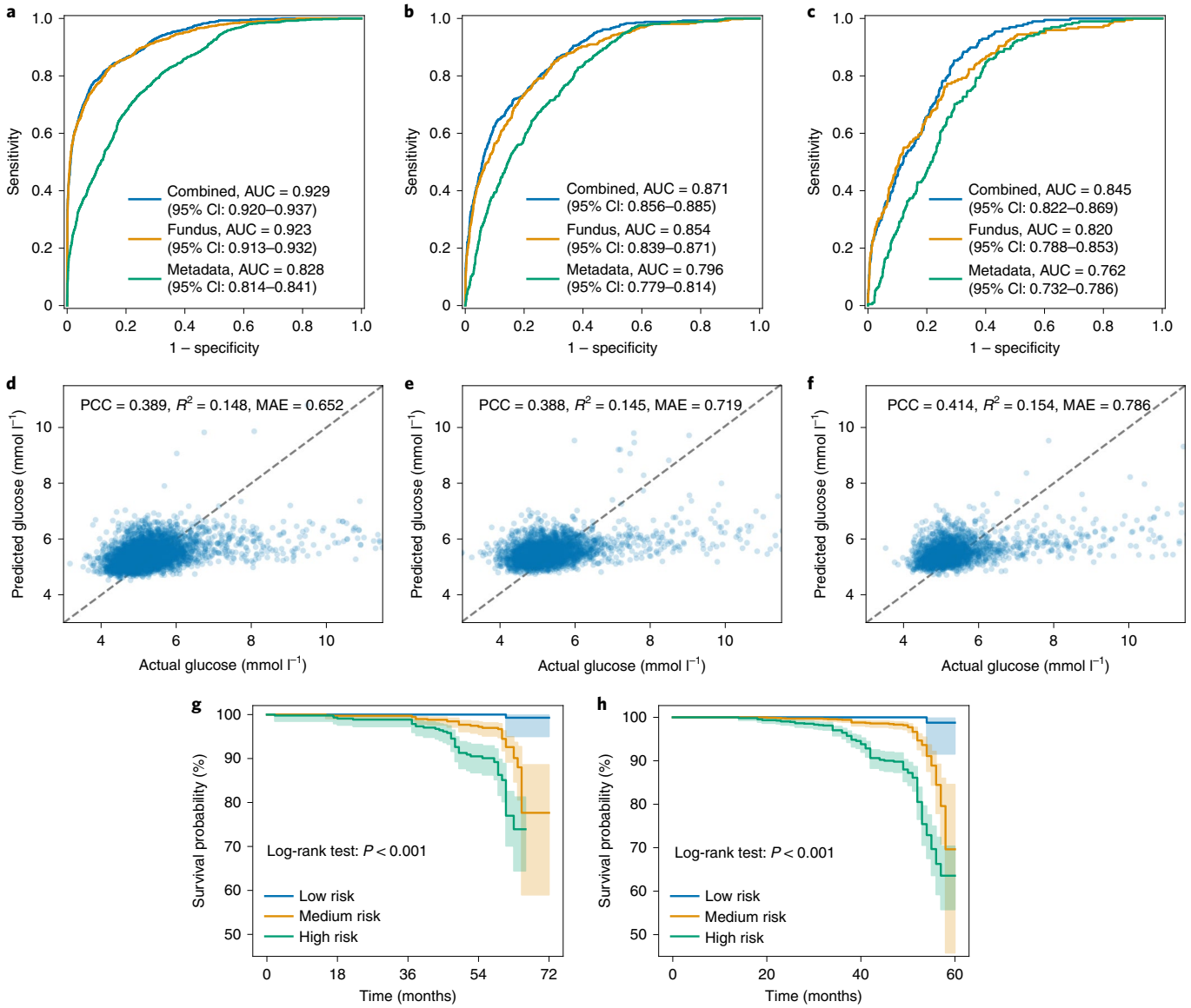

Fig. 5 | Performance of the AI system in the identification and incidence prediction of T2DM. a–c, ROC curves showing performance of T2DM detection using the metadata-only model, the fundus-only model and the combined model on: the internal test set (case:control = 2,361:6,366) (**a**); external test set 1 (case:control = 2,823:5,236) (**b**); and external test set 2: point-of-care study (case:control = 1,189:1,892) (**c**). **d–f**, Correlation analysis of predicted blood glucose versus actual blood glucose on: the internal test set (**d**), external test set 1 (**e**) and external test set 2: point-of-care study (**f**). **g**,**h**, Kaplan–Meier plots showing the incidence of T2DM during follow-up visits in the internal longitudinal test set (**g**) and external longitudinal test set (**h**). Survival curves represent the high-risk, medium-risk and low-risk subgroups and the shaded regions represent the 95% CI.

complications of diabetes are related to vascular damage, retinal fundus images provide highly relevant data for non-invasive and longitudinal assessment of vascular health. Although fasting blood glucose testing is relatively straightforward, non-invasive fundus image-based screening could broaden access and enable early detection and lifestyle modifications. In terms of monitoring disease progression, compared with HbA1c and random blood glucose testing, image-based analysis of vascular changes in the retina could provide alternative and objective means of assessing diabetic control and tissue damage.

For real-world clinical applications, the operating point of an AI system could be set differently to balance the positive predictive value (PPV) and the negative predictive value (NPV). Performance metrics for CKD and T2DM detection in the cohorts were determined by the operating points selected from the tuning dataset (shown in Supplementary Tables 8 and 9). We generated specific values of sensitivity, specificity, PPV and NPV of the AI system for systemic disease detection in order to meet various screening needs or clinical applications. To identify CKD cases with high confidence, we used a very-high decision threshold with a relatively higher PPV. For the external test set, we achieved a PPV of 88.4% (95% CI: 83.9–92.8) and 89.3% (95% CI: 80.8–95.5), while retaining a sensitivity of 34.8% (95% CI: 31.8–38.6) and 29.9% (95% CI: 21.5–34.3) (Supplementary Table 8). To identify healthy controls from CKD cases with high confidence, we used a very-low decision threshold with a relatively higher NPV. For the external test set, we achieved a NPV of 99.7% (95% CI: 99.5–99.9) and 99.4% (95% CI: 98.5–100.0), while retaining a specificity of 37.5% (95% CI: 36.1–38.9) and 32.1% (95% CI: 29.0–35.5) (Supplementary Table 8).

This study has assessed the role of automated AI algorithms in the detection of CKD and T2DM using a low-cost smartphone-based imaging device. Camera-enabled smartphones are widely available

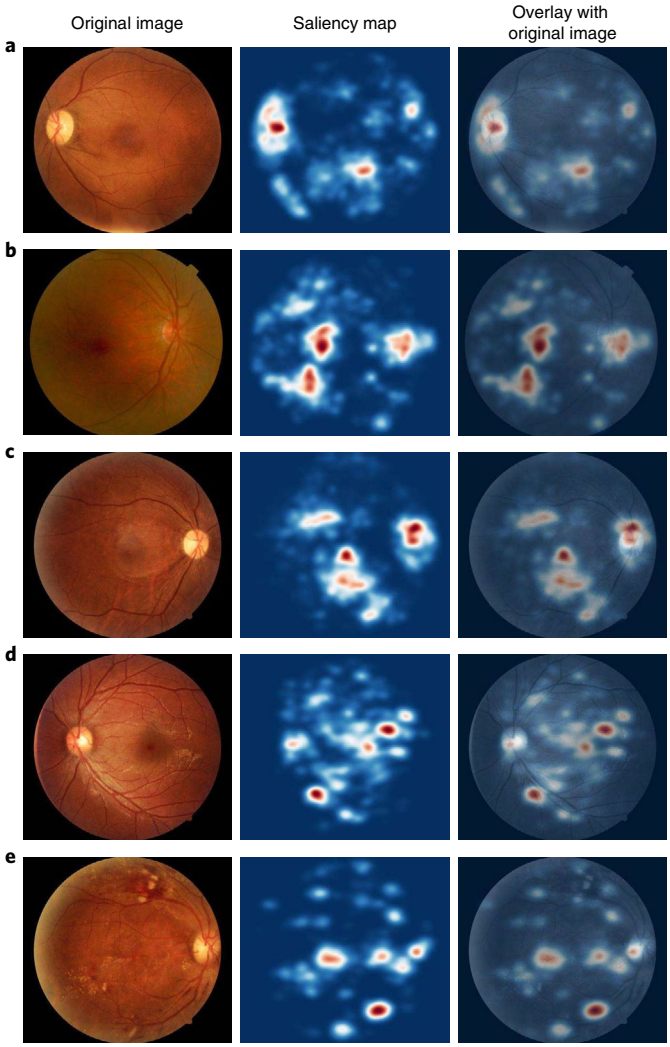

Fig. 6 | Gradient visualizations of AI predictions of CKD staging and T2DM using the integrated gradient algorithm. Visual explanations of the areas of the images most important for the determination of the model prediction for qualitative review and clinical relevance. The columns are (1) the original fundus image, (2) a saliency map and (3) a saliency map overlaying the original image. **a**–**e**, Early CKD (corresponding to stage 1 and 2) (**a**), advanced CKD (corresponding to stage 3) (**b**), severe+ CKD (corresponding to stage 4+) (**c**), T2DM, with 'no signs of diabetic retinopathy' (NDR) (**d**) and T2DM, with DR (**e**).

around the world. The availability of retina specialists and trained retinal graders is, however, a major limitation in most countries. The portability and easy operation of our approach helps to provide screening in remote areas without the involvement of health care professionals. Therefore, the deployment of a smartphone- and cloud-based AI fundus diagnosis system could broaden access irrespective of regional resource variations and improve early detection of these treatable systemic diseases. In addition, this approach has potential as a non-invasive screening for multiple diseases in the general population and is not limited to patients with diabetes, as ocular imaging becomes more widely available. Such systems would also lead to significant cost savings over traditional disease screening programmes.

This study has several differences to consider in comparison to previous studies[20,21]. Earlier literature in this area utilized a different definition of CKD. We used a more widely adopted definition of CKD based on the consensus clinical guideline: an eGFR $\geq$60 ml min^{-1} per 1.73 m^2 with albuminuria or eGFR <60 ml min^{-1} per 1.73 m^2, while in a previous study for predicting CKD, the authors used a narrower clinical definition of CKD as a binary outcome on the basis of eGFR <60 ml min^{-1} per 1.73 m^2. Second, our time-to-critical-event model based on longitudinal cohorts could provide great utility in managing patients during their early disease course. Knowing which patients will progress to CKD or advanced+ stages from their first retinal fundus images will permit providers to triage and manage these patients while optimizing resources appropriately.

Moreover, as systemic diseases usually affect both eyes equally, we predicted with the AI system at the image level and averaged the image-level output at the patient level for a systemic condition prediction. To confirm consistency between two eyes, we further conducted experiments comparing AI prediction performance for CKD based on data from the left eye only and right eye only. This comparison showed a very strong inter-eye correlation (Supplementary Fig. 3f), supporting the accuracy and validity of the AI model. To investigate whether there is a bias in the study design on detecting an undiagnosed disease, we subdivided the documented CKD and T2DM patients into those with and without prior diagnoses of CKD and T2DM at the time of health screening or image capture (Supplementary Fig. 10). Together, the results suggest that the history of clinically diagnosed or undiagnosed CKD and T2DM at the time of health screening does not significantly affect the performance of our fundus image-based AI models.

Our study has several limitations which we hope to address in the future. First, since our AI was trained in a solely Chinese population and tested in both an external Chinese cohort from several different geographic areas (external test sets 1 and 2), its generalizability in other racial populations will need to be further validated. Therefore, we added another external multi-ethnicity validation cohort (Supplementary Table 10) consisting of individuals from Kashi (also called Kashgar) in Xinjiang Autonomous Region (Uighur ethnicity) and Macau (Portuguese ethnicity), in which the AI model showed good performance (Supplementary Fig. 6). Additional training on more diverse clinical and demographic cohorts may further improve diagnostic accuracy and clinical utility in a broad range of populations. Another limitation is related to the measurement of eGFR; inaccurate estimation may exist in patients with rapid deterioration of renal function. In this study, patients with combined acute kidney injury (AKI) diagnosis (or a low incidence of AKI) were excluded to minimize this interference. Thus, such a calculation strategy can still reflect the characterization of the patient's renal function progression. In addition, whether additional clinical metadata (such as blood pressure trends, smoking status, alcohol consumption level and family history) could further improve the accuracy of the predictions needs to be explored. While the model appears well suited for diagnostic screening, it remains limited in its ability to provide prognostic information to individual patients or insights into pathogenic mechanisms based on saliency maps.

Methods

Dataset characteristics. To develop an AI system for the detection of CKD and T2DM, fundus image data were collected from the CC-FII, which included the following participants: COACS in Tangshan City, Hebei Province; the Peking University First Affiliated Hospital and the Peking University Third Affiliated Hospital, both in Beijing; West China Hospital in Chengdu, Sichuan Province; Renji Hospital in Chongqing; Kunshan Hospital of Jiangsu University, Kunshan, Jiangsu Province; and Zhong Shan Ophthalmic Center of Sun Yat-sen University and Nanfang Hospital both in Guangzhou, Guangdong Province. All procedures were performed as a part of patients' routine annual health checks. Institutional review board and ethics committee approvals were obtained in all locations and all participating patients signed a consent form.

The COACS is a community-based, prospective study to investigate how suboptimal health status contributes to the incidence of non-communicable chronic diseases in Chinese adults[32]. This COACS study has two phases: a cross-sectional survey followed by a longitudinal study. The participants were

recruited from Tangshan city, which is a large, modern industrial city and adjoins two mega cities: Beijing and Tianjin. In phase I, all participants underwent clinical laboratory measurements. In the second phase, a long-term yearly clinical follow-up has been performed until 2024, with the purpose of better understanding how suboptimal health, environmental and genetic risk factors contribute to the development of major chronic diseases. We have elected to use this cohort for our study because it balances healthy participants and those with major chronic diseases such as CKD and T2DM.

For CKD and T2DM detection, we used retinal fundus images from retrospective datasets. The first one, a cross-sectional cohort (CC-FII-C) from CC-FII, included a total of 86,312 fundus images from 43,156 participants as the developmental dataset of our AI models for systemic disease detection. All participants from the developmental cohort were split into mutually exclusive sets for a training set (70%) and a tuning set (10%), as well as an internal test set (20%). Detailed participant characteristics are shown in Table 1. External test set 1 included participants undergoing an annual health check in Zhong Shan Ophthalmic Center of Sun Yat-sen University and Nanfang Hospital both in Guangzhou, Guangdong Province. Study participants also subsequently underwent ophthalmological examinations with fundus imaging. Only retinal fundus images were included in this study with 8,059 participants from external test set 1. We further tested our AI system in external test set 2, which is a prospective point-of-care setting, for the evaluation of the generalizability of the AI system with smartphone-based devices. For this prospective study, we enroled a total of 3,081 participants, including 228 patients with CKD and 1,189 patients with T2DM in Tangshan from 16 September 2019 to 15 November 2019. These participants were part of the COACS study, yet they were independent of the development dataset. To demonstrate the generalizability of the AI model in other ethnic groups, we added another external multi-ethnicity validation cohort consisting of 615 individuals from the First Regional Hospital of Kashi in Xinjiang Autonomous Region as well as the University Hospital of Macau University of Science and Technology and Hospital Kiang Wood, both in Macau, China.

To develop an AI system for the prediction of the incidence of CKD and T2DM, we further used two longitudinal cohorts (Supplementary Table 1). The first longitudinal dataset CC-FII-L is a subset from CC-FII, which consisted of 10,269 participants for routine annual health checks with a follow-up period of six years. All the participants from the CC-FII-L were randomly divided into a developmental set and an internal longitudinal test set (3,376 individuals) in an 8:2 ratio. Another longitudinal dataset, an external longitudinal test set, was used for external validation of the AI system for incidence prediction of the systemic diseases. This is a population-based study of Chinese from Beijing, China, which recruited patients from hospitals or health centres for annual health checks, including DR screening in Peking University's First Affiliated Hospital and Third Affiliated Hospital. A total of 3,376 participants were used as external validation (external longitudinal test set) of our AI models for the incidence prediction of CKD and T2DM. The patient characteristics for each cohort can be found in Supplementary Table 1.

Image quality control. The retinal fundus images were captured using a variety of standard fundus cameras, including Topcon TRC-NW6 (Topcon), Zeiss Visucam 224 (Carl Zeiss Meditec AG), Canon CR6–45NM (Canon) and KOWA Nonmyd α-DIII (Kowa). During the image grading process, all fundus images were first de-identified to remove any patient-related information. Participants were recruited from an annual health check population that underwent a physical examination by a physician and fundus image photography. For screening and grading retinal fundus images of DR, a hierarchical two-tier grading process was performed by ten phase I and five phase II graders. Phase I graders consisted of individuals trained by ophthalmologists and evaluated to perform at least 95% accuracy determined by a quiz consisting of 1,000 fundus images of various retinal diseases. Phase II graders consisted of ophthalmologists who individually reviewed every image classified by phase I graders. To check consistency among phase II graders, 20% of images were randomly selected and reviewed by three senior retinal specialists. The second tier of five ophthalmologists independently read and verified the true labels for each image. To account for disagreement, the evaluation test set was also checked by expert consensus. About 9% of the study participants were excluded due to poor photographic quality or unreadable images and missing clinical diagnosis. After establishing the consensus diagnoses, images were transferred to the AI team to develop a deep-learning algorithm for image-based classification.

Definition and criteria for disease diagnosis. The following criteria were used to define systemic disease and related disease staging. According to the international kidney disease clinical classification based on the Kidney Disease: Improving Global Outcomes (KDIGO) Clinical Practice Guideline, the staging and actual risk of adverse outcomes of kidney disease are stratified by the renal glomerular filtration function defined as the GFR categories[33]. GFR is an indicator of overall kidney function, which equals the total amount of fluid filtered through the functioning nephrons per unit of time[34]. The eGFR in this study was based on the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) Equation. This equation has been extensively validated in Chinese and Asian populations[35,36].

In clinical practice, CKD was diagnosed as an eGFR of more than 60 ml min^{-1} per 1.73 m^2 with albuminuria or less than 60 ml min^{-1} per 1.73 m^2, confirmed in at least two visits separated by three months. Healthy controls were defined as eGFR above 60 ml min^{-1} per 1.73 m^2 without albuminuria, checked by a negative urine dip-stick test. In our study, we utilized the images and corresponding eGFR measurements of patients who had already been diagnosed with CKD. For patients with multiple visits, we use the labels of retinal fundus images taken during the visit when the diagnosis was first established.

Once a diagnosis of CKD has been made, the next step is to determine staging, which is determined by the extent to which eGFR has decreased. Staging of kidney function is classified with eGFR as follows: stage 1 (eGFR ≥90 ml min^{-1} per 1.73 m^2 with albuminuria), stage 2 (eGFR 60–89 ml min^{-1} per 1.73 m^2 with albuminuria), stage 3 (eGFR 30–59 ml min^{-1} per 1.73 m^2), and stage 4+ (eGFR <30 ml min^{-1} per 1.73 m^2). For the CKD stage classification, we defined potential or mild kidney impairment, called early CKD, as stages 1 and 2. Stage 3 is denoted as advanced CKD. Additionally, one of the important goals of CKD management is to prevent end-stage kidney failure, which is associated with the need for renal dialysis or transplantation and with increased mortality. Detection and early intervention in severe+ CKD are also crucial, which is defined with an eGFR cut-off of 30 ml min^{-1} per 1.73 m^2 (corresponding to stage 4 or above).

T2DM was diagnosed by fasting blood glucose ≥7.0 mmol l^{-1} at least two times, an HbA1c value of 6.5% or more and/or a history of drug treatment for diabetes. For the detection of T2DM patients, there exist two subsets of fundus images: a DR group and a NDR group. In this study, we conducted an experiment of NDR versus healthy controls to investigate the performance of detecting asymptomatic T2DM patients. We also compared the performance with/without DR, and the results show that the AI system could learn early eye features of T2DM patients and showed comparable results with/without DR. In our detection study, we utilized the images and corresponding clinical data, including laboratory values, of patients who had already been diagnosed with T2DM.

Categories of tasks for systemic disease detection. To develop a diagnostic capable of detecting vascular-related systemic diseases based on retinal images, we trained separate deep-learning models for each task. More specifically, we had two types of prediction task: regression and classification. For regression tasks, we trained two models to predict continuous values of eGFR and fasting blood glucose, respectively. The detection of systemic disease or staging was treated as a classification task. We also performed binary classification for T2DM detection. For each of these tasks, including CKD detection and T2DM detection, we compared the performance of three prediction models, each with a different set of input data. As a baseline, we used metadata-only models., which comprised age, sex, blood pressure, hypertension, height, weight and body-mass index (BMI), to develop random-forest and logistic-regression classifiers. We included diabetes as a covariate when building our AI model in CKD prediction. We further used fundus-only models based on deep CNNs with fundus images. Finally, we trained combined models that integrate fundus image data and clinical information. The image feature vector derived from the CNN model was concatenated with the clinical feature of the same patient. A multilayer perception (MLP) took these features as input for classification.

For training and tuning the random-forest model, the number of trees in the forest 'n_estimators' was with the default setting, and the maximum number of leaf nodes 'max_leaf_nodes' and the minimum number of samples required to be at a leaf node 'min_samples_leaf' were tuned by parameter search. Finally, they were set with n_estimators = 100, max_leaf_nodes = 31 and min_samples_leaf = 20. Similarly, we added logistic regression models for CKD and T2DM detection, with regularized likelihood estimation (L2 penalty) and regularization coefficient $C = 1.0$ (Supplementary Table 2).

Fundus image enhancement. To capture the non-specific anatomical and physiological features on fundus images relevant to systemic diseases, we proposed image enhancement to improve the performance of the AI models. Two methods were utilized for fundus image enhancement, including Contrast Limited Adaptive Histogram Equalization (CLAHE)[37] and colour normalization[38]. CLAHE enhancement is conducted by dividing the image into local regions and applying histogram equalization over all neighbourhood pixels. Specifically, we converted the input fundus images from RGB colour space (red, green and blue channels) into LAB colour space (lightness, green–red opponent, and blue–yellow opponent channels). After applying the CLAHE on the lightness channel, we then converted it back to RGB. Compared with the original fundus images, CLAHE algorithm enhanced the details and visibility level of the fundus image. The fundus image normalization method was performed as follows: $x' = \alpha x - \alpha\text{Gauss}(x,\mu,\Sigma,s \times s) + \beta$, where x is the input image, x' is the normalized image, α and β are parameters, and $\text{Gauss}(x,\mu,\Sigma,s \times s)$ is the Gaussian filtered image with a Gaussian kernel (μ,Σ) of size (μ,Σ). We used $\alpha = 4$ and $\beta = 128$, $\Sigma = I$ and $s = 10$, following the settings of Liu et al.[38]. With image normalization, we could reduce the colour scope bias among fundus images taken under different lighting or device conditions. The effectiveness of the image enhancement method can be seen in Supplementary Fig. 1.

Model development and training. CNNs were used to analyse and classify the fundus images in this study. With ResNet-50[39] as the backbone, we pretrained on the ImageNet dataset for all deep-learning models demonstrated. ResNet-50 is a five-stage network with one convolution and four identity blocks, which utilizes skip connections to overcome the degradation problem of deep-learning models. For regression tasks of continuous values prediction (eGFR and fast blood glucose), a fully connected layer with one scalar as output was used as the final layer in the ResNet-50 model. For binary classification tasks, an additional softmax layer beside a fully connected layer was attached to the model. Retraining consisted of loading the convolutional layers with pretrained weights, newly initializing additional layers for our regression and binary classification tasks and training models on the corresponding development sets.

We used a three-layer MLP for combined models. Each of the two hidden layers has 128 nodes and was applied with the ReLU activation function. The MLP was jointly trained with the CNN. The MSE loss was used as an objective function for the regression tasks of prediction of continuous values and the binary cross entropy loss was used for binary classification tasks. Training of models by back-propagation of errors was performed in batches of 32 images resized to 512×512 pixels for 50 epochs with a learning rate of 10^{-3}. Training was performed using the Adam optimizer with a weight decay of 10^{-6}. Transformations of random horizontal and vertical flip were added to each batch during training as data augmentation in order to enable an improved and generalized network learning. The models were implemented using PyTorch. We randomly divided the developmental dataset into a training set (7/8 of the development set) and a tuning set (1/8 of the development set) to develop our model. The models selected for evaluation on validation sets were the models with the best validation loss on the tuning set. There were no samples overlapping at the patient level in training and evaluation sets.

Model ensemble. To improve the overall performance of the AI, we applied a model ensemble. For each task, we trained four model instances with different processed fundus images as input. Each input image was pre-processed into three variations by applying CLAHE only, normalization only, and both CLAHE and normalization. Thus, for each task, we had four models with the same architecture trained in parallel on the same development set but with each using differently pre-processed fundus images. The reported predictions were obtained by averaging the outputs of the four model instances.

Prediction of the development of systemic diseases using longitudinal cohorts. For the incidence analysis of the CKD, we denoted the index data as the time when the participants were without CKD at baseline. All participants who underwent a urine analysis with a negative result at the baseline visit using urine dip-stick test were included for the CKD incidence analysis. The incidence data were denoted as the time when the participants were recorded as having CKD or advanced+ CKD during the follow-up visits. Similarly, we predicted the development of T2DM, with the index data defined as one without T2DM at the first visit. The development of T2DM was diagnosed as T2DM incidence data (or end-point) within the yearly clinical follow-up.

We trained the CPH models on the training and tuning set using variables based on the metadata and fundus image-based risk score. The metadata-based model comprised age, sex, blood pressure, height, weight, BMI, hypertension and T2DM. The image-based risk core is the predicted z-score (standard score) of the first visit generated from the CKD/T2DM detection model and used to predict progression risks of patients in combination with metadata. According to the risk scores of the first visit from the CPH model for the CKD/T2DM detection, the patients are triaged into three groups: low, medium, and high risk according to the upper and lower quartiles of predicted risk scores in the tuning set, respectively. Supplementary Fig. 11 shows the distribution of the risk scores and the related thresholds (the upper and lower quartiles) across datasets. The risk scores were also treated as categorical variables according to quartiles during the incidence analysis on validation sets (Table 2).

Kaplan–Meier curves were constructed for the risk groups, and the significance of differences between group curves was computed using the log-rank test. Time-dependent ROC curves[40] were used to quantify model performance on validation sets at the time of interest. ROC curves were constructed at a landmark time from predicted risk scores of relative patients made using the model. The univariable and multivariable CPH models were fitted. Two multivariable CPH models were developed, a combined metadata and fundus model and a metadata-only model serving as a baseline model. Statistical significance of HRs and adjusted HRs of CPH models were evaluated using the likelihood ratio test.

Interpretation of AI predictions. The difficulty in obtaining clinically intelligible features remains the greatest drawback of artificial neural network-based systems. A visualization tool is needed that would enable clinicians to understand important clinical variables in real time. To this end, we employed the integrated gradient algorithm[41] to produce visual explanations. Gradient-based visualization methods use gradients to quantify the importance of each pixel in the image. The importance of each pixel in the image to the correct predictions of the models can be quantified by $\frac{\partial y_i}{\partial x}$, where y_i is the model output for class i and x is the input image. However, gradient saturation presents a problem where the gradients of the output with respect to the input can be small even though they are important for the model output. Such phenomena can happen when the model outputs for the correct class reach a certain magnitude. To overcome this problem, the integrated gradient method improves the measurement of importance as follows: $I_i(x) = (x - x') \times \int_{\alpha=0}^{1} \frac{\partial f(x' + \alpha(x - x'))}{\partial x} \mathrm{d}\alpha$, where I_i is the integrated gradients for pixel i, x is the input image, x' is the baseline image, x_i and x' are values of pixel i in x and x', f is the model to be visualized. The saliency maps generated by integrated gradient indicate the effect of each pixel on the model predictions. We applied Gaussian filtering to saliency maps for smoothness on the fundus images.

Statistical analysis. To evaluate the performance of regression models for continuous values prediction in this study, we calculated MAE, R^2 and PCC. We applied the Bland–Altman plot[42] to display the difference between the measured eGFR and the predicted value of a sample against the average of the two. With 95% limits of agreement and ICC, we evaluated the agreement of the predicted eGFR and actual eGFR. We calculated the ratio between the variance of the model outputs and the variance of ground-truthed data using the tuning set to calibrate outputs. The models' performance on binary classification predictions was evaluated by ROC curves of sensitivity versus 1 – specificity. The AUC of ROC curves were reported with 95% CIs. The 95% CIs of AUCs were estimated with the non-parametric bootstrap method (1,000 random resampling with replacement). Sensitivity and specificity were determined by the selected thresholds on the validation set. The prediction of continuous values of eGFR and blood glucose level were evaluated with regression models. CKD and T2DM detection were evaluated with binary classification models. For each patient, we made a prediction with the AI system at the image level and then averaged the image-level output at a patient level for a final prediction in each patient. ICC was used to assess the agreement between AI predicted values of left and right eyes, where stage CKD was measured by predicted eGFRs and diabetes, measured by log-likelihood ratios of predicted probability of T2DM presence. We calculated the incidence rate for the whole cohort and for each risk group (three strata) as the number of events per 1,000 person-years at risk. The Byar Poisson approximation method was used to calculate 95% CIs of incidence[43]. Then Kaplan–Meier estimators were constructed for different risk groups, and the significance of differences between groups was tested by log-rank tests. CPH models were tested using the likelihood ratio test. We used the time-dependent AUC at four years and five years to measure model performance. The Kaplan–Meier curve and the time-dependent ROC-AUC were calculated using the Python packages of lifelines (version 0.25.5) and scikit-survival (version 0.14.0).

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

Data availability

Restrictions apply to the availability of the developmental and validation datasets, which were used with permission of the participants for the current study. De-identified data may be available for research purposes from the corresponding authors on reasonable request.

Code availability

The deep-learning models were developed and deployed using standard model libraries and the PyTorch framework. The models can be trained via the publicly available ResNet-50 architecture starting from the pretrained models, available at https://github.com/pytorch/vision. Custom codes were specific to our development environment and used primarily for data input/output and parallelization across computers and graphics processors. The codes may be available for research purposes from the corresponding authors on reasonable request.

Received: 13 November 2020; Accepted: 12 May 2021;
Published online: 15 June 2021

References

1. GBD Chronic Kidney Disease Collaboration. Global, regional, and national burden of chronic kidney disease, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* **395**, 709–733 (2020).
2. Levin, A. et al. Global kidney health 2017 and beyond: a roadmap for closing gaps in care, research, and policy. *Lancet* **390**, 1888–1917 (2017).
3. Kooman, J. P., Kotanko, P., Schols, A. M., Shiels, P. G. & Stenvinkel, P. Chronic kidney disease and premature ageing. *Nat. Rev. Nephrol.* **10**, 732–742 (2014).
4. Saeedi, P. et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: results from the International Diabetes Federation Diabetes Atlas, 9(th) edition. *Diabetes Res. Clin. Pract.* **157**, 107843 (2019).
5. Wong, T. Y. & Sabanayagam, C. The war on diabetic retinopathy: where are we now. *Asia Pac. J. Ophthalmol.* **8**, 448–456 (2019).

6. Balakumar, P., Maung, U. K. & Jagadeesh, G. Prevalence and prevention of cardiovascular disease and diabetes mellitus. *Pharmacol. Res.* **113**, 600–609 (2016).
7. From the Center of Disease Control and Prevention. Lower extremity amputation episodes among persons with diabetes–New Mexico, 2000. *JAMA* **289**, 1502–1503 (2003).
8. American Diabetes Association. 11. Microvascular complications and foot care: standards of medical care in diabetes-2020. *Diabetes Care* **43**, S135–S151 (2020).
9. Luk, A. O. et al. Quality of care in patients with diabetic kidney disease in Asia: The Joint Asia Diabetes Evaluation (JADE) Registry. *Diabet. Med.* **33**, 1230–1239 (2016).
10. Wu, B., Zhang, S., Lin, H. & Mou, S. Prevention of renal failure in Chinese patients with newly diagnosed type 2 diabetes: a cost-effectiveness analysis. *J. Diabetes Investig.* **9**, 152–161 (2018).
11. Esteva, A. et al. A guide to deep learning in healthcare. *Nat. Med.* **25**, 24–29 (2019).
12. Cheung, C. Y., Tang, F., Ting, D. S. W., Tan, G. S. W. & Wong, T. Y. Artificial intelligence in diabetic eye disease screening. *Asia Pac. J. Ophthalmol.* **8**, 158–164 (2019).
13. Ravizza, S. et al. Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data. *Nat. Med.* **25**, 57–59 (2019).
14. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
15. Wang, K., Liu, X., Zhang, K., Chen, T. & Wang, G. Anterior segment eye lesion segmentation with advanced fusion strategies and auxiliary tasks. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020* **12265**, 656–664 (Springer, 2020).
16. Kermany, D. S. et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**, 1122–1131(2018).
17. Liang, H. et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nat. Med.* **25**, 433–438 (2019).
18. Wang, G. et al. A deep-learning pipeline for the diagnosis and discrimination of viral, non-viral and COVID-19 pneumonia from chest X-ray images. *Nat. Biomed. Eng.* https://doi.org/10.1038/s41551-021-00704-1 (2021).
19. Poplin, R. et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat. Biomed. Eng.* **2**, 158–164 (2018).
20. Rim, T. H. et al. Prediction of systemic biomarkers from retinal photographs: development and validation of deep-learning algorithms. *Lancet Digit. Health* **2**, e526–e536 (2020).
21. Sabanayagam, C. et al. A deep learning algorithm to detect chronic kidney disease from retinal photographs in community-based populations. *Lancet Digit. Health* **2**, e295–e302 (2020).
22. Liu, T. Y. A. Smartphone-based, artificial intelligence-enabled diabetic retinopathy screening. *JAMA Ophthalmol.* **137**, 1188–1189 (2019).
23. Chen, C., Lee, G. G., Sritapan, V. & Lin, C. Deep convolutional neural network on iOS mobile devices. In *2016 IEEE International Workshop on Signal Processing Systems (SiPS)* 130–135 (IEEE, 2016); https://doi.org/10.1109/SiPS.2016.31
24. Wu, Y., Lim, J. & Yang, M. H. Object Tracking Benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**, 1834–1848 (2015).
25. Schroff, F., Kalenichenko, D. & Philbin, J. FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 815–823 (IEEE, 2015); https://doi.org/10.1109/CVPR.2015.7298682
26. Vos, T. et al. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet* **388**, 1545–1602 (2016).
27. Gansevoort, R. T. et al. Lower estimated GFR and higher albuminuria are associated with adverse kidney outcomes. A collaborative meta-analysis of general and high-risk population cohorts. *Kidney Int.* **80**, 93–104 (2011).
28. Levey, A. S. & Coresh, J. Chronic kidney disease. *Lancet* **379**, 165–180 (2012).
29. Group, E. T. D. R. S. R. Grading diabetic retinopathy from stereoscopic color fundus photographs—an extension of the modified Airlie House classification: ETDRS report number 10. *Ophthalmology* **98**, 786–806 (1991).
30. Tuot, D. S. et al. Chronic kidney disease awareness among individuals with clinical markers of kidney dysfunction. *Clin. J. Am. Soc. Nephrol.***6**, 1838–1844 (2011).
31. Tuttle, K. R. et al. Diabetic kidney disease: a report from an ADA Consensus Conference. *Am. J. Kidney Dis.* **64**, 510–533 (2014).
32. Wang, Y. et al. China suboptimal health cohort study: rationale, design and baseline characteristics. *J. Transl. Med.* **14**, 291 (2016).
33. Levin, A. et al. Kidney disease: Improving global outcomes (KDIGO) CKD work group. KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney Int. Suppl.* **3**, 1–150 (2013).
34. Levey, A. S., Becker, C. & Inker, L. A. Glomerular filtration rate and albuminuria for detection and staging of acute and chronic kidney disease in adults: a systematic review. *JAMA* **313**, 837–846 (2015).
35. Bikbov, B. et al. Global, regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* **395**, 709–733 (2020).
36. Liao, Y., Liao, W., Liu, J., Xu, G. & Zeng, R. Assessment of the CKD-EPI equation to estimate glomerular filtration rate in adults from a Chinese CKD population. *J. Int. Med. Res.* **39**, 2273–2280 (2011).
37. Pisano, E. D. et al. Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms. *J. Digital Imaging* **11**, 193–200 (1998).
38. Liu, P. et al. Large-scale left and right eye classification in retinal images. *Comput. Pathol. Ophthalmic Med. Image Anal.* **11039**, 261–268 (2018).
39. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (2016); https://doi.org/10.1109/CVPR.2016.90
40. Kamarudin, A. N., Cox, T. & Kolamunnage-Donà, R. Time-dependent ROC curve analysis in medical research: current methods and applications. *BMC Med. Res. Method.* **17**, 53 (2017).
41. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. In *Proc. of the 34th International Conference on Machine Learning-Volume 70* **3319** (2017).
42. Giavarina, D. Understanding Bland Altman analysis. *Biochemia Med.* **25**, 141–151 (2015).
43. Breslow, N. & Day, N. *Statistical Methods in Cancer Research. Volume II–The Design and Analysis of Cohort Studies* **82**, 1–406 (IARC Scientific Publications, 1987).

Acknowledgements

This study was funded by the National Natural Science Foundation of China (61906105, 61872218 and 61721003), National Key Research and Development Program of China (2019YFB1404804, 2017YFC1104600, 2017YFC0112402), 1.3.5 project for disciplines of excellence, West China Hospital, Sichuan University (ZYJC20001, ZYJC18010), Guangzhou Regenerative Medicine and Health Guangdong Laboratory, Macau University of Science and Technology, Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua University Initiative Scientific Research Program, and Guoqiang Institute, Tsinghua University, Wellcome Trust (216593/Z/19/Z). We thank members of the Zhang, Yuan and Wang groups for their assistance. We thank many volunteers and physicians for grading retinal photographs.

Author contributions

K. Zhang, X.L., J.X., J.Y., W.C., K.W., T.C., Y.G., S.N., X.X., X.Q., Y. Su, W.X., A.O., K.X., Z.L., M.Z., X. Zeng, C.Z., O.L., E.Z., J.Z., Y.X., D.K., K. Zhou, Y.P., S.L., I.L., Y.C., C.W., M.P., G.Z, Q.Z., J.L., D.L., X. Zou, A.W., J.W., Y. Shen, F.F.H., P.Z., T.X., Y.Z. and G.W. collected and analysed the data. K. Zhang and G.W. conceived and supervised the project. K. Zhang and G.W. wrote the manuscript with assistance from K.X. All authors discussed the results and reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41551-021-00745-6.

Correspondence and requests for materials should be addressed to K. Zhang, T.C., T.X., Y.Z. or G.W.

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

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

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

nature research

Corresponding author(s): Kang Zhang

Last updated by author(s): May 9, 2021

Reporting Summary

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

Statistics

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

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

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No special software or code was used to collect the data.
Data analysis	Pytorch and python libraries.

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

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Restrictions apply to the availability of the developmental and validation datasets, which were used with permission of the participants for the current study. De-identified data may be available for research purposes from the corresponding authors on reasonable request.

Field-specific reporting

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

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

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

Sample size	We developed AI models capable of identifying chronic kidney disease (CKD) and type 2 diabetes mellitus (T2DM) using a total of 115,344 retinal fundus photographs from 57,672 patients.
Data exclusions	No data were excluded if they have passed the initial image-quality-control step.
Replication	Replication was not relevant. We used independent validation cohorts.
Randomization	Samples were randomly allocated to the training, tuning and testing sets.
Blinding	During image processing, all images were first de-identified to remove any patient related information.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Retinal fundus images and metadata were obtained as a part of routine clinical care and a prospective study.
Recruitment	Participants were recruited from multiple hospitals.
Ethics oversight	The China Consortium of Fundus Image Investigation (CC-FII) Ethics Committee.

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