An AI system for the detection and incidence prediction of chronic kidney disease and type 2 diabetes based on retinal fundus images

Kang Zhang[1*#], Xiaohong Liu[2*], Jin Yuan[3*], Jie Xu[4*], Wenjia Cai[3*], Kai Wang[5], Ting Chen[2#], Yuanxu Gao[1], Sheng Nie[6], Xiaoqi Qin[5], Wenqin Xu[1], Andrea Olvera[1], Kanmin Xue[7], Zhihuan Li[1], Yuandong Su[8], Meixia Zhang[8], Charlotte L. Zhang[9], Oulan Li[9], Edward E. Zhang[9], Jie Zhu[10], Yiming Xu[2], Daniel Kermany[1], Kaixin Zhou[9], Ying Pan[10], Shaoyun Li[12], Iat Fan Lai[13], Ying Chi[14], Changuang Wang[15], Qi Zhang[16], Johnson Lau[17], Dennis Lam[18], Yin Shen[19], Tao Xu[9#], Yong Zhou[20], and Guangyu Wang[5#]

[1] *Center for Biomedicine and Innovations, Faculty of Medicine, Macau University of Science and Technology, Macau, China*
[2] *Department of Computer Science and Technology, Tsinghua University, Beijing, China*
[3] *State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China*
[4] *Beijing Institute of Ophthalmology, Beijing Tongren Eye Center, Beijing Tongren Hospital, Beijing Ophthalmology and Visual Science Key Lab, Beijing, China*
[5] *School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China.*
[6] *State Key Laboratory of Organ Failure Research; National Clinical Research Center for Kidney Disease & Nanfang hospital, Southern Medical University, Guangzhou, China*
[7] *Nuffield Department of Neuroscience, Oxford University, Oxford, UK*
[8] *Center for Translational Innovations, West China Hospital and Sichuan University, Chengdu, China*
[9] *Bioland Laboratory, Guangzhou, China*
[10] *Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, China*
[11] *Department of Endocrinology, Kunshan Hospital Affiliated to Jiangsu University, Kunshan, China*
[12] *The Big Data Research Center, Chongqing Renji affiliated Hospital to the University of Chinese Academy of Sciences, Chongqing, China.*
[13] *Ophthalmic Center, Kiang Wood Hospital, Macau, China*
[14] *Peking University First Affiliated Hospital, Beijing, China*
[15] *Peking University Third Affiliated Hospital, Beijing, China*
[16] *Biotherapy Center, Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China*
[17] *Department of Applied Biology and Chemical Technology, Hong Kong Polytechnic University, Hong Kong, China*
[18] *C-MER Dennis Lam and Partners Eye Center, C-MER International Eye Care Group, Hong Kong, China*
[19] *Medical Research Institute, Wuhan University, Wuhan, China*
[20] *Clinical Research Center, Shanghai First People's Hospital, Shanghai Jiaotong University, Shanghai, China*

[#] These authors contributed equally to this work.
*Correspondence and reprints should be addressed to: guangyu.wang24@gmail.com; tingchen@mail.tsinghua.edu.cn; xutao@ibp.ac.cn; or kang.zhang@gmail.com

Editorial corresponding author:

Kang Zhang, MD, PhD
Center for Biomedicine and Innovations, Faculty of Medicine
Macau University of Science and Technology
Email address: kang.zhang@gmail.com

Abstract

Regular health screening plays a crucial role in the early detection of common chronic diseases and prevention of their progression. An AI system capable of recapitulating early disease detection, staging and incidence prediction would help to improve healthcare access and delivery, particularly in resource-poor or remote settings. Using a total of 115,344 retinal fundus photographs from 57,672 patients (with data split into mutually exclusive training, internal testing, and external validation sets), we first developed AI models capable of identifying chronic kidney disease (CKD) and type 2 diabetes mellitus (T2DM) based on fundus images. The AI system was shown to be capable of predicting the clinical indicators of CKD and T2DM (including eGFR and blood glucose levels), which indicates its potential for extracting quantitative clinical metrics embedded subtly within retinal fundus images. We further developed an AI system to predict the risk of disease progression using baseline images of 10,269 patients for whom longitudinal clinical data were available for up to 6 years, which demonstrated potential utility in optimizing health screening intervals and clinical management. The generalizability of the AI system in identifying and predicting the progression of CKD and T2DM was evaluated using population-based external validation cohorts. Moreover, a prospective pilot study with 3,081 patients was also conducted to demonstrate the broader applicability of the AI system at the 'point-of-care' using fundus images captured with smartphones. The results provide proof-of-concept for a reliable and non-invasive AI-based clinical screening tool based on fundus photographs for the early detection and incidence prediction of two common systemic diseases.

Introduction

Systemic diseases, including chronic kidney disease (CKD) and diabetes mellitus, pose major healthcare 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 outcome is not limited to end-stage renal failure requiring dialysis or transplantation but also 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 Center 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 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 diabetic patients with over 5-year disease course to an annual test on urinary albumin and estimated glomerular filtration rate (eGFR). A previous study reported a less than 20% awareness rate of CKD and a less than 50% treatment rate[9] in China. 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 carrying out kidney disease screening in newly diagnosed T2DM patients[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 homeostasis of the body, in which we can 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 healthcare[11-15], including image-based diagnosis[16] and natural language processing[17]. In particular, convolutional neural networks 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.

Though there is a recent report on advanced CKD detection using a retinal image-based AI system[21], prediction of CKD onset from a normal baseline and diagnosis of early CKD (proteinuria) have not yet been described and would have an important impact on disease prevention and outcomes. Similarly, predicting the onset of type 2 diabetes 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 novel possibility of better channeling healthcare 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 healthcare, 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 platforms 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 iPhone X (Apple Inc, Cupertino, CA, USA)[23], smartphones now deploy a range of machine learning algorithms from image classification[24] to object tracking[25], face detection and recognition[26], natural language translation[27], sentence completion[28], and augmented reality[29]. Furthermore, smartphone ownership is widespread worldwide. For instance, in Nigeria, the physician-to-patient ratio is 1: 2660[30], while smartphone ownership is 1:3.5[31]. An AI-based smartphone diagnostic platform is thus an attractive way to broaden healthcare access by encouraging patients to self-monitor and allowing doctors to diagnose and follow-up patients remotely.

In this study, we aimed to develop an AI system capable of analyzing retinal fundus images to detect CKD and T2DM. We employed deep learning-based analysis of retinal fundus images for two types of tasks: 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 (**Fig.1a**). 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. 1b**). Using this approach, targeted screening to subgroups of the population could potentially help 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' platform for CKD and T2DM screening in the community.

Results

Definitions of chronic kidney disease and type 2 diabetes

Based on international guidelines and previous studies[1,32], identifying individuals with CKD relies primarily on estimated glomerular filtration rate (eGFR), an index of kidney function, and renal damage markers (e.g. urinary albumin). The presence of CKD was defined by an eGFR ≥60 mL/min/1.73m^2 with albuminuria or eGFR <60 mL/min/1.73 m^2, confirmed in at least two visits separated by three months; whereas normal controls were defined by an eGFR ≥60 ml/min/1.73m^2 without albuminuria. In our study, we utilized the images and corresponding eGFR measurements of already-diagnosed CKD patients.

Furthermore, CKD can be divided into 5 stages depending on severity during clinical applications. For the purpose of this study, we defined three risk stratifications depending on the 5 stages: early CKD (stage 1 and 2, eGFR ≥60 mL/min/1.73m^2 with albuminuria); advanced CKD (stage 3, eGFR 30-59 mL/min/1.73m^2), and severe+ CDK (stage 4 and 5, eGFR <30 mL/min/1.73m^2). Early identification of early CKD to advanced CKD is clinically important as a timely intervention to modify known risk factors (e.g. treatment of hypertension, diabetes and urinary tract obstructions) could slow disease progression and reduce long-term health and economic burdens. 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 increased mortality[33].

For the study of the other systemic disease of T2DM, we utilized the images and corresponding clinical data, including laboratory values of already-diagnosed T2DM patients, based on: fasting blood glucose level ≥7.0 mmol/L confirmed in at least two visits, or an HbA1c value ≥6.5%, and/or based on a history of drug treatment for diabetes. In addition, 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[34].

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 **Extended Data 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 are subsets from the CC-FII (**Fig. 1a and Table 1**). All subjects from the CC-FII-C were split randomly into mutually exclusive sets for training, tuning and "internal testing" of the AI algorithm at a 70%:10%:20% 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,** see more details in **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 healthcare professional or a patient to capture fundus images for assessment (**Extended Data Fig. 10**). The images could be uploaded to a Health Insurance Portability and Accountability Act (HIPAA)-compliant cloud service where the AI platform could autonomously grade incoming diagnostic requests. In this prospective study from September to November 2019, 3,081 patients were recruited in the COACS cohort with 6,162 smartphone-captured fundus images (see **Methods**).

To predict the development of CKD and T2DM, we used de-identified fundus images from two longitudinal datasets (**Extended Data Table 1**, see more details in **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 **Extended Data 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; see more details in **Methods**). Based on the definitions of CKD, we trained AI models to perform binary classification tasks. We first developed two models: baseline random

forest models using clinical metadata; and deep learning models using fundus images. We hypothesized that a model utilizing both clinical metadata and fundus images might be even more accurate, thus developing 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% 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 AUC was 0.842 (95% CI: 0.827-0.856) for the clinical metadata-only model, 0.885 (95% 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)**.

Early CKD is defined by eGFR >60 mL/min/1.73m^2 with albuminuria (a sign of kidney damage), whereas normal controls have eGFR＞60 mL/min/1.73m^2 without albuminuria. AI-based prediction of early CKD from normal 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 and 2e)**. 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 the 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 performances based on fundus images alone. For the detection of CKD, the AI platform delivered an AUC of 0.817 (95% CI: 0.785-0.842) of the metadata-only model, an AUC of 0.870 (95% CI: 0.847-0.893) of the fundus-only model, and 0.897 (95% CI: 0.855-0.902) of the combined model in the external test set 2: "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 an AUC 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 metadata-only, fundus-only, and combined model 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

platform administered via personal mobile devices. This platform could assist medical professionals in screening and monitoring systemic microvascular diseases.

GFR prediction and CKD staging using retinal fundus images

Estimated GFR (eGFR) is normally calculated based on measurement of serum creatinine level (see more details in Methods). We tested whether a patient's eGFR from measured serum creatinine could be predicted by an AI model using their fundus images alone. 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) and mean absolute error (MAE) of 11.08 (95% CI: 10.80-11.35) (**Fig. 3a**). The model performed similarly well on the first external test set with an ICC of 0.62 (95% CI: 0.60-0.64) and an MAE of 12.91 (95% CI: 12.58-13.24) (**Fig. 3b**). 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) and an MAE of 11.85 (95% CI: 11.42-12.27) (**Fig. 3c**). Bland-Altman plots showed a negative proportional bias (slope of linear fit); that is, the AI-based model underestimated the GFR level to a greater extent at high levels of GFR than at low levels. This proportional bias (slope of linear fit) was reduced after the model outputs were calibrated to have the same variance as that of the ground-truth measurements (Methods) (**Extended Data Fig. 3a-3c**).

Furthermore, a linear correlation between the AI-predicted GFR and measured eGFR showed strong associations with a coefficient of determination (R2) of 0.507 and Pearson's correlation coefficient (PCC) of 0.716 for the internal test set 1 (from CC-FII) (**Fig. 3d**). When evaluated on the external validation cohorts, the AI model achieved R2 of 0.481 and PCC of 0.700, and R2 of 0.327 and PCC of 0.577, for external test set 1 and 2, respectively (**Fig. 3e** and **3f**). 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.

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 tasks 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) in the internal validation dataset A (**Extend Data 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) (**Extend Data 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 (**Extend Data Fig. 3e**).

Prediction of the development of CKD using longitudinal cohorts

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

The performance of progression prediction models of CKD/advanced CKD using the Cox proportional hazards (CPH) model has been summarized in **Extended Data Table 4**. The metadata-based model achieved a C-index of 0.756 (0.699-0.810) on the internal test set and a C-index of 0.651 (0.569-0.730) on the external test set. When the deep learning features extracted from fundus images were combined with clinical metadata, the model performance improved to a C-index of 0.845 (0.789-0.910) on the internal test set and 0.719 (0.627-0.807) on the external test set. The combined progression prediction model had a statistically significant improvement in comparison to clinical metadata only-based prediction, as shown in **Extended Data Table 4** (Permutation test). In addition, we conducted analysis on the univariable and multivariable hazard ratio (HR) comparing the fundus-based models to known risk factors (**Extended Data Table 3)**. We used HRs to examine and qualify the influence of specific factors on the rate of occurrence of a particular event rate (e.g., onset of disease) at a particular point in time. As shown in Extended Table 3, the adjusted HR of the fundus predictor is significant ($p<0.001$).

Using the 6-year longitudinal data on CKD staging from our developmental patient cohorts (CC-FII-L), we further used the Kaplan-Meier method to stratify normal patients into three risk groups (the low, medium or high risks) for developing CKD or advanced to severe CKD (advanced+ CKD). The incidence of the CKD (per 1000 person-years) stratified by risk groups of the AI model is shown in **Table 2**. For the Kaplan-Meier curves and log-rank tests, thresholds for the high-risk and the low-risk 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 and 4c**). To assess the generalizability of the AI model, the same tests were performed on an external longitudinal test set of 3,215 patients with 5 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 and 4d)**. The Kaplan-Meier curves for the risk stratification from the metadata-only model are illustrated in **Extended Data Fig.7. Extended Data 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 ($p<0.001$).

The prognostic accuracy of the AI system was assessed using time-dependent ROC analysis. The AUC for predicting the development of CKD was 0.844 (95% CI: 0.787-0.888) in the internal test set. We further used the AUC at 4 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 (**Extended Data Fig. 9a** and **9b**).

Using the AI system to identify T2DM

To further 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 (**Extended Data Table 1**).

We first evaluated our models' performance in the detection of T2DM patients from normal 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 in the internal test set **(Fig. 5a**). When evaluated on the first external test set, the model performed well, with an AUC 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 an AUC 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**).

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: (i) T2DM patients with diabetic retinopathy (DR) and (ii) T2DM patients with 'No Diabetic Retinopathy' (NDR) (**Extended Data Fig. 5a** and **5b**). 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 diabetic retinopathy. These results indicate that the AI model could detect T2DM based on fundus images prior to any apparent clinical manifestations of diabetic retinopathy (**Extended Data Fig. 5c**).

We further tested the ability of our models to predict the level of mean blood glucose from fundus images alone. Interestingly, the AI model achieved a relatively strong performance in the internal test set, the external test set 1 and the external test set 2, suggesting that blood glucose levels could be predicted and quantified from fundus images alone (**Extended Data Fig. 4a-4c**). Bland-Altman analysis and relating calibration were also performed to investigate the agreement between two approaches of the blood glucose measurement (**Extended Data Fig. 4d-4i,** see more details in **Methods**).

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 normal patients over a 5-year period. The normal patients within the developmental cohort were stratified into the 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 1000 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. 5c** and **5d**).

Subsequently, we performed progression analysis of the T2DM using CPH models. The performance of the metadata-based model and combined model is summarized in **Extended Data Table 5**. The metadata-based model achieved a C-index of 0.774 (0.732-0.819) on the internal test set and a C-index of 0.746 (0.706-0.775) on the external test set. When the deep learning features extracted from fundus images were combined with clinical metadata, the model performance improved to a C-index of 0.781(0.743-0.819) on the internal test set and 0.765 (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 (**Extended Data Table 3).**

Among a total of 3,064 subjects with verified disease development outcomes, 185 (6.0%) developed T2DM within a 5-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)

in the internal longitudinal test set and 0.824 (95%CI: 0.793-0.858) in the external longitudinal test set (**Extended Data Fig. 9c and 9d**).

Visualization of evidence for CKD/T2DM prediction

Finally, to improve the interpretability of the AI model and shed light on its diagnostic mechanism, Integrated Gradients (IG) 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** and **Extended Data Fig. 5**. While microvascular pathologic changes are known to exist in CKD/T2DM, they may not be observable in fundus photographs, at least not to an ophthalmologist's eyes. 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 (e.g. in cases of diabetic retinopathy). 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 vessel branch points, arterial-venous junctions, and/or exudates or haemorrhages, suggesting that vascular health may play a role (**Fig. 6**). 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 (**Extended Data Fig. 5a** and **5b**).

Discussion

The common systemic microvascular diseases, CKD and T2DM, are major public health burdens and early detection is critical for successful clinical management and favorable outcomes. We hypothesized that the AI model might be able to diagnose CKD and T2DM by isolating and identifying subclinical changes that are undetectable by human observers. In this study, we developed an AI platform to detect CKD and T2DM with high degrees of accuracy and predict future disease development based on fundus images alone and prior to any clinical manifestation. These AI-derived diagnostic capabilities were not only applicable to clinical-grade retinal images obtained using professional cameras but also applicable 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 the 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[35]. In this regard, a recent report on CKD diagnosis using fundus images raises new possibilities for early disease detection[21]. However, 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 the quality-of-life. Given the great 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 capture of patients for early intervention. In this regard, the AI system should be viewed as a welcome addition to assist physicians and better direct healthcare 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[36]. Our retinal fundus image-based AI system showed its value for T2DM screening in general clinics and communities. As the majority of complications of diabetes are related to vascular damage, retinal fundus images provide highly relevant data for non-invasive and longitudinal assessment of vascular health. Though fasting blood glucose testing is relatively straightforward, the non-invasive fundus image-based screening system could broaden access and enable early detection and life-style modifications. In terms of monitoring for disease progression, compared with the 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 true positive rate (TPR) and the false-positive rate (FPR). Performance metrics for CKD/T2DM detection in the cohorts were determined by the operating points selected from the tuning dataset (shown in **Extended Data Table 6 and 7**). 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% (83.9-92.8) and 89.3% (80.8-95.5), while retaining a sensitivity of 34.8% (31.8-38.6) and 29.9% (21.5-34.3) (**Extended Data Table 6**). To identify normal controls with high confidence, we used a very-low decision threshold with a relatively higher NPV. For the external test set, we achieved a PPV of 99.7% (99.5-99.9) and 99.4% (98.5-100.0), while retaining a specificity of 37.5% (36.1-38.9) and 32.1% (29.0-35.5) (**Extended Data Table 6**).

In addition, this study has assessed the role of automated AI algorithms in the detection of CKD/T2DM using a low-cost smartphone-based imaging device. Camera-enabled smartphones are widely available around the world. The availability of retina specialists/trained retinal graders is, however, a major limitation in most countries. The portable and easy operation of our approach helps to provide screening in remote areas without the involvement of healthcare professionals. Therefore, the deployment of a smartphone-cloud based AI fundus diagnosis platform could provide access to essential diagnosis 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 programs.

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 ≥60 mL/min/1.73m^2 with albuminuria or eGFR <60 mL/min/1.73m^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 per 1·73 m². In contrast, in this study, CKD was defined by an eGFR ≥60 mL/min/1.73m^2 with albuminuria, or eGFR <60 mL/min, based on the clinical guidelines. Secondly, our time-to-critical event model based on longitudinal cohorts could provide a great utility in managing patients during their early disease course. Knowing which patients will progress to a CKD or advanced+ stages from a patient's 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 based on the left eye only and right eye only data for CKD detection. This comparison showed a very strong inter-eye correlation (see more details in **Extended Data Fig. 3f)**, supporting the accuracy and validity of the AI model.

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 (the external test set 1&2), its generalizability in other racial populations will need to be further validated. Therefore, we added another external multi-ethnicity validation cohort (**Extended Data Table 8**) 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 (**Extended Data 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 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 need 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 China Consortium of Fundus Image Investigation (CC-FII), which included the following participants: China suboptimal health cohort study (COACS) in Tangshan City, Hebei Province; the Peking University Frist-Affiliated Hospital and the Peking University Second-Affiliated Hospital, both in Beijing; West China Hospital in Chengdu, Chongqing Renji Hospital in Chongqing, both in Sichuan Province; Kunshan Hospital of Jiangshu University, Kunshan, Jiangshu Province. All procedures were performed as a part of a patients' routine annual health check. Institutional Review Board (IRB)/Ethics Committee approvals were obtained in all locations and all participating subjects signed a consent form.

The COACS is a community-based, prospective study to investigate how suboptimal health status contributes to the incidence of NCD (non-communicable chronic diseases) in Chinese adults[37]. 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 subjects and those with major chronic diseases such as CKD and T2DM.

For the 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 subjects as the developmental dataset of

our AI models for systemic disease detection. All subjects from the developmental cohort were split into mutually exclusive sets for a training set and a tuning set (70%:10%), as well as an internal test set (20%). Detailed participant characteristics are shown in **Table 1**. The external test set 1 included subjects 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 the external test set 1. We further tested our AI system in the external test set 2, which is a prospective "point-of-care" settings, for the evaluation of the generalizability of the AI system with smartphone-based device. For this prospective study, we enrolled a total of 3,081 subjects, including 228 CKD patients, and 1,189 T2DM patients in Tangshan from September 16, 2019, to November 15, 2019. These participants were a 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 400 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 (**Extended 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 6 years. All the subjects from the CC-FII-L were randomly divided for developmental set and 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 centers for annual health checks, including diabetic retinopathy screening in Peking University's First Affiliated Hospital and Third Affiliated Hospital. A total of 3,376 subjects 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 **Extended Data Table 1**.

Image quality control

The retinal fundus images were captured using a variety of standard fundus cameras, including Topcon TRC-NW6 (Topcon Corp, Tokyo, Japan), Zeiss Visucam 224 (Carl Zeiss Meditec AG, Jena, Germany), Canon CR6-45NM (Canon Inc, Tokyo, Japan), and KOWA Nonmyd α-DIII (Kowa Company Ltd, Aichi, Japan). 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 diabetic retinopathy, 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 1000-fundus-image quiz 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 3 senior retinal specialists. The second tier of five ophthalmologists independently read and verified the true labels for each image. In order 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/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 the systemic disease and related disease staging. According to the international kidney disease clinical classification based on the KDIGO guideline (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 glomerular filtration rate (GFR) categories[38]. GFR is an indicator of overall kidney function, which equals the total amount of fluid filtered through the functioning nephrons per unit of time[39]. The estimated GFR (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[40,41].

In clinical practice, CKD was diagnosed as an eGFR of more than 60 mL/min/1.73 m² with albuminuria or less than 60 mL/min/1.73 m², confirmed in at least two visits separated by three months. Once a diagnosis of CKD has been made, the next step is to determine staging, which is determined by the extent to which GFR has decreased. Staging of kidney function is classified with eGFR as follows: Stage 1 (eGFR＞90 mL/min/1.73 m², with albuminuria), Stage 2 (eGFR 60-89 mL/min/1.73 m² with albuminuria), Stage 3 (eGFR 30-59 mL/min/1.73 m²), and Stage 4+ (eGFR<30 mL/min/1.73 m²). For the CKD stage classification, we defined the potential or mild kidney impairment, called early CKD, corresponding to Stages 1 and 2. Stage 3 is denoted as advanced CKD. In addition, the detection and early intervention of CKD severe+ are also crucial, which is defined with an eGFR cutoff with 30 mL/min/1.73 m^2 (corresponding to stage 4 or above). Normal controls were defined as eGFR above 60 ml/min/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 already-diagnosed CKD patients. For a patient with multiple visits, we use the labels

of retinal fundus images taken corresponding to the visit when first establishing the diagnosis.

T2DM was diagnosed by fasting blood glucose ≥7.0 mmol/L at least 2 times, or as an HbA1c value of 6.5 % or more, and/or by a history of drug treatment for diabetes. For the detection of T2DM patients, there exist two subsets of fundus images: a diabetic retinopathy (DR) group and a 'diabetes mellitus with no DR' group (NDR) group. In the study, we conducted an experiment of NDR versus normal 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 already-diagnosed T2DM patients.

Deep learning and transfer learning methods

Categories of tasks for systemic diseases detection

To develop a diagnostic platform which is 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 tasks, including a "regression" task and a "classification" task. For the 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 the 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 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 convolutional neural networks (CNNs) with fundus images. Finally, we trained combined models which 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 multi-layer 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 the CKD and T2DM detection, with regularized likelihood estimation (L2 penalty) and regularization coefficient C=1.0 (**Extended Data 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)[42] and color normalization[43]. CLAHE enhancement is conducted by dividing the image into local regions and applying histogram equalization over all neighborhood pixels. Specifically, we converted the input fundus images from RGB color space into LAB color space. 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: $x' = \alpha x - \alpha Gauss(x, \mu, \Sigma, s \times s) + \beta$, where x is the input image, x' is the normalized image, α and β are parameters, and $Gauss(x, \mu, \Sigma, s \times s)$ is the Gaussian filtered image with a Gaussian kernel (μ, Σ) of size $s \times s$. We used $\alpha = 4$ and $\beta = 128$, $\Sigma = I$, and $s = 10$, following the settings of Liu et al[43]. With image normalization, we could reduce the color scope bias among fundus images taken under different lighting or device conditions. The effectiveness of the image enhancement method can be seen in **Extended Data Fig. 1**.

Model development and training

Convolutional neural networks (CNNs) were used to analyze and classify the fundus images in this study. With the ResNet-50[44] 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 besides 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 Mean-Square Error (MSE) loss was used as an objective function for the "regression" tasks of prediction of continuous values and the Binary Cross Entropy (BCE) 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[45]. 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 subjects 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 subjects were recorded as having CKD/ 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 Cox' proportional hazard (CPH) models on the training and tunning set
using variables based on the metadata and fundus image, which comprised age, sex, blood pressure, height, weight, BMI, hypertension, T2DM, and predicted z-score (standard score) of the first visit generated from the CKD/T2DM detection model.

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. 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[46] 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 hazard ratios and adjusted hazard ratios of CPH models were evaluated using the likelihood ratio test.

Handheld fundoscopy using a smartphone attachment

A 22-diopter double-convex aspheric condensing lens (Volk Optics, Ohio, USA) was mounted inside a custom-designed smartphone fundoscope attachment (**Extended Data Fig. 10**). The plastic components of the fundoscope were computer-designed on SolidWorks with collision simulation, converted to Standard Triangle Language format, and 3D printed (fused filament fabrication) in polylactic acid to a resolution of 100 microns. The lens was stably anchored to the fundoscope cone using a rubber ring and printed lens locking system. Two perpendicular polarizing filters were also incorporated within the optical system to minimize the reflection of the smartphone flashlight from the cornea.

We calculated the cone-shaped offset of the iPhone (Apple Inc, USA) from the condensing lens and the distance between the condensing lens and the anterior principal plane of the eye using 12.3 cm as the focal length of the condensing lens. Thus, the condensing lens worked in harmony with the iPhone camera light source and the optical system of the eye to achieve Maxwellian illumination of the retina and project an in-focus and widefield image onto the camera sensor. Fundus images were captured using this smartphone-mounted fundoscope in a prospective study within the COACS study. Informed consent was obtained from patients prior to pupil dilation and retinal photography using the standard operating procedure below. The same AI system for detecting CKD or T2DM as used for analyzing professional fundus camera-derived images was used for the handheld fundoscopy-derived images, which detected systemic diseases of CKD or T2DM. The performance of the model was evaluated using ROC curves.

Imaging protocol using the smartphone attachment

Standard operating procedure for fundus image capture using the smartphone fundoscope attachment:

1. The pupil is dilated with a drop of 1% tropicamide.
2. Select ‘New Patient’ from the main program display and enter patient information.
3. First, hold the iPhone X (Apple Inc, Cupertino, CA, USA) with fundoscope attachment in the right hand approximately 10 cm from the patient’s eye to obtain a red reflex at the center of the display.
4. Then, slowly move the imaging device towards the eye while keeping the red reflex centered on the display. When the red reflex fills the entire field-of-view, stabilize the end of the fundoscope attachment with the left hand, which leans on the patient’s forehead for stability.

5. Make fine adjustments to the distance between the end of the fundoscope attachment and the eye to obtain a focused image of the retina. Press the capture button on the iPhone to acquire the image.
6. Base on the quality of images obtained, the photographer may acquire up to 9 images per eye, with the best quality image being selected for the study.

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[47] in order 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 improve the measurement of importance as follows: $I_i(x) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial f(x' + \alpha(x - x'))}{\partial x_i} d\alpha$, where I_i is the integrated gradients for pixel i, x is the input image, x' is the baseline image, x_i and x_i' 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 Mean Absolute Error (MAE), R-square (R2), and Pearson Correlation Coefficient (PCC). We applied the Bland-Altman plot[48] 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 Intraclass Correlation Coefficient (ICC), we evaluated the agreement of the predicted eGFR and actual eGFR. The models' performance on binary classification predictions was evaluated by Receiver Operating Characteristic (ROC) curves of sensitivity versus 1 – specificity. The Area Under the Curve (AUC) of ROC curves were reported with 95% Confidence Intervals (CIs). The 95% CIs of AUC 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. The CKD and T2DM detection were evaluated with binary classification models. For each subject, 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 incidence rate for the whole cohort and for each risk group by three-strata shown 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[49]. Then Kaplan-Meier estimators were constructed for different risk groups, and the significance of differences between groups was tested by log-rank tests. We used the time-dependent AUC (area under the curve) 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).

Data availability statement

To protect patient confidentiality, we have deposited patient retinal image data in a secured and patient confidentiality compliant cloud in China in concordance with data security regulations. Data access can be requested by writing to the corresponding authors. All data and code access requests will be reviewed and (if successful) granted by the Data Access Committee.

Code availability statement

The deep learning models were developed and deployed using standard model libraries and the PyTorch framework. Custom codes were specific to our development environment and used primarily for data input/output and parallelization across computers and graphics processors. We will make our codes available upon request and approval by a Data Access Committee.

Acknowledgements

This study was funded by the National Key Research and Development Program of China (2019YFB1404804, 2017YFC1104600, 2017YFC0112402), Guangzhou Regenerative Medicine and Health Guangdong Laboratory, National Natural Science Foundation of China (61906105, 61872218 and 61721003), Macau University of Science and Technology, Tsinghua University Initiative Scientific Research Program. We thank members of Zhang, Yuan, and Wang groups for their assistance. We thank many volunteers and physicians for grading retinal photographs.

Author Contributions

J.X., J.Y., ZH.L., WJ. C., WQ. Xu., X. L., A. O., GX. Z., LH. Z., C. Z., OL. L., E. Z., J. Z., SQ. H., KM. X., W. W., S. N., M. L., J. Z., M. P. V., M. A., JQ. W., A. W., XG. Z., Q. Z., P.T., KX.Z., T.X., GY. W., and K. Z. collected and analyzed the data. K.Z. and GY.W. conceived and supervised the project. J.Y-N.L and D. L. help with the

data analysis and interpretation. K.Z., KM.X., GY.W. wrote the manuscript. All authors discussed the results and reviewed the manuscript.

Competing Interests statement

The authors declare no competing financial interests.

References

1. Collaboration, G.B.D.C.K.D. 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. *Nature reviews. Nephrology* **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 research and clinical practice* **157**, 107843 (2019).
5. Leasher, J.L.*, et al.* Global Estimates on the Number of People Blind or Visually Impaired by Diabetic Retinopathy: A Meta-analysis From 1990 to 2010. *Diabetes care* **39**, 1643-1649 (2016).
6. Balakumar, P., Maung, U.K. & Jagadeesh, G. Prevalence and prevention of cardiovascular disease and diabetes mellitus. *Pharmacological research* **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, A. 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. *Diabetic medicine : a journal of the British Diabetic Association* **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. *Journal of diabetes investigation* **9**, 152-161 (2018).
11. Esteva, A.*, et al.* A guide to deep learning in healthcare. *Nature medicine* **25**, 24-29 (2019).
12. Norgeot, B., Glicksberg, B.S. & Butte, A.J. A call for deep-learning healthcare. *Nature medicine* **25**, 14-15 (2019).
13. Ravizza, S.*, et al.* Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data. *Nature medicine* **25**, 57-59 (2019).
14. Topol, E.J. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine* **25**, 44-56 (2019).

15. De Fauw, J., *et al.* Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine* **24**, 1342-1350 (2018).
16. Kermany, D.S., *et al.* Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell* **172**, 1122-1131 e1129 (2018).
17. Liang, H., *et al.* Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nature medicine* **25**, 433-438 (2019).
18. Wang, C., Elazab, A., Wu, J. & Hu, Q. Lung nodule classification using deep feature fusion in chest radiography. *Computerized Medical Imaging and Graphics* **57, 10-18**(2017).
19. Poplin, R., *et al.* Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature biomedical engineering* **2**, 158-164 (2018).
20. Rim, T.H., *et al.* Prediction of systemic biomarkers from retinal photographs: development and validation of deep-learning algorithms. *The Lancet Digital 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. *The Lancet Digital Health* **2**(2020).
22. Liu, T.Y.A. Smartphone-Based, Artificial Intelligence–Enabled Diabetic Retinopathy Screening. *JAMA Ophthalmology* **137**, 1188-1189 (2019).
23. Chen, C.-F., Lee, G.G., Sritapan, V. & Lin, C.-Y. in 2016 IEEE International Workshop on Signal Processing Systems (SiPS). *130-135 (IEEE)* (2016).
24. Howard, A.G., *et al.* MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv preprint arXiv* **1704.04861**(2017).
25. Wu, Y., Lim, J. & Yang, M.H. Object Tracking Benchmark. *IEEE transactions on pattern analysis and machine intelligence* **37**, 1834-1848 (2015).
26. Schroff, F., Kalenichenko, D. & Philbin, J. in Proceedings of the IEEE conference on computer vision and pattern recognition. **815-823**.
27. Bahdanau, D., Cho, K. & Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *ArXiv* **1409**(2014).
28. Mikolov, T., Chen, K., Corrado, G.s. & Dean, J. Efficient Estimation of Word Representations in Vector Space. *Proceedings of Workshop at ICLR* **2013**(2013).
29. Ignatov, A., *et al. AI Benchmark: Running Deep Neural Networks on Android Smartphones*, (2018).
30. Organization, W.H. Density of physicians (total number per 1000 population, latest available year). *Last accessed Oct* **http://www. who. int/gho/health_workforce/physicians_density/en.** (2016).
31. Poushter, J. Smartphone ownership and internet usage continues to climb in emerging economies. *Pew Research Center* **22, 1-44** (2016).
32. 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. *The Lancet* **388**, 1545-1602 (2016).
33. Gansevoort, R.T., *et al.* Lower estimated GFR and higher albuminuria are associated with adverse kidney outcomes in both general and high-risk populations. *Kidney International Supplements* **80(1):93-104**(2011).

34. 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).
35. Tuot, D.S.*, et al.* Chronic kidney disease awareness among individuals with clinical markers of kidney dysfunction. *Clinical journal of the American Society of Nephrology : CJASN* **6**, 1838-1844 (2011).
36. Tuttle, K.R.*, et al.* Diabetic kidney disease: a report from an ADA Consensus Conference. *American journal of kidney diseases : the official journal of the National Kidney Foundation* **64**, 510-533 (2014).
37. Wang, Y.*, et al.* China suboptimal health cohort study: rationale, design and baseline characteristics. *Journal of translational medicine* **14**, 291 (2016).
38. 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 International Supplements* **3**, 1-150 (2013).
39. 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).
40. 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. *The Lancet* **395**, 709-733 (2020).
41. 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. *Journal of International Medical Research* **39**, 2273-2280 (2011).
42. Pisano, E.D.*, et al.* Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms. *Journal of digital imaging* **11**, 193-200 (1998).
43. Liu, P.*, et al.* Large-Scale Left and Right Eye Classification in Retinal Images. in *Computational Pathology and Ophthalmic Medical Image Analysis* (eds. Stoyanov, D.*, et al.*) 261-268 (Springer International Publishing, Cham, 2018).
44. He, K., Zhang, X., Ren, S. & Sun, J. *Deep Residual Learning for Image Recognition*, (2016).
45. Paszke, A.*, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library*, (2019).
46. Kamarudin, A.N., Cox, T. & Kolamunnage-Donà, R. Time-dependent ROC curve analysis in medical research: Current methods and applications. *BMC Medical Research Methodology* **17**(2017).
47. Sundararajan, Mukund, Taly, A. & Yan, Q. Axiomatic Attribution for Deep Networks. *In Proceedings of the 34th International Conference on Machine Learning-Volume 70* **3319**(2017).
48. D, G. Understanding bland altman analysis. *Biochemia medica* **25(2): 141-151**(2015).
49. Breslow, N. & Day, N. Statistical Methods in Cancer Research. Volume II--The Design and Analysis of Cohort Studies. *IARC scientific publications*, 1-406 (1987).

Figure Legend

Figure 1. Schematic illustration of an AI system for detection and incidence prediction of systemic diseases using retinal fundus images. a, Model development of the AI system. The system made two different types of predictions: "regression tasks" for continuous values (eGFR and fasting blood glucose); binary "classification" tasks for categorical values (CKD detection and T2DM). The AI prediction is generated with an ensemble of model instances. **b,** Application and evaluation of the AI system. The left panel: Training on a developmental dataset for assessment of CKD/CKD staging/ T2DM. The model was then tested on an independent cohort to ensure the generalizability. The middle panel: we developed an AI platform to predict future disease development based on fundus images from longitudinal cohorts, which should help strengthen CKD/T2DM surveillance. The right panel: a prospective study on a "Point of care" setting using a smartphone. Fundus images captured on the phone were transmitted to a cloud hosting the AI model which produces an instant report transmitted back to the smartphone, an ophthalmologist and a hospital. CKD, chronic kidney disease; T2DM, type 2 diabetes mellitus.

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

Figure 3. Model performance in assessing GFR from retinal fundus images. a-c, Bland-Altman plot for the agreement between the predicted and estimated GFR. **The** X-axis represents the mean of predicted and estimated GFR, and the Y-axis represents the difference between the two measurements. Assessing agreement between predicted and estimated GFR with mean absolute error (MAE) and intraclass correlation coefficient (ICC)**. a,** In internal test set. **b,** In external test 1. **c,** In external test 2, a prospective point-of-care pilot study.
d-f, Correlation analysis of the predicted eGFR vs actual eGFR generated using the regression model, with mean absolute error (MAE), Pearson's correlation coefficient (PCC) and R squared (R2). **d,** AI performance on internal test set. **e,** AI performance on the external test set 1. **f,** AI performance on the external test set 2: "point-of-care" study.

Figure 4. Kaplan-Meier plots illustrating 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 colors are the high-risk, medium-risk and low-risk subgroups stratified by the upper and lower quartiles in the tuning dataset. **a and b,** The incidence of CKD in **a,** the internal longitudinal test set and **b,** in the external longitudinal test set. **c and d,** The incidence of advanced+ CKD (corresponding to stage 3 or more severe) in **c,** internal longitudinal test set and **d,** in the external longitudinal test set. P-value is obtained from the log-rank test.

Figure 5. Performance in identifying and incidence prediction of T2DM with the AI models. a and b, AI performance in detecting T2DM using the metadata-only model, the fundus-only model and the combined model. **a**, ROC curves showing performance of T2DM detection on internal test set (case: control=2,361:6,366). **b**, ROC curves showing performance of T2DM detection on external test set 1(case: control=2,823:5,236). **c**, ROC curves showing performance of T2DM detection on external test set 2: "point-of-care" study (case: control=1,189:1,892). **d and e,** Kaplan-Meier plots showing the incidence of T2DM during the follow-up visits in **c,** internal longitudinal test set and **d,** external longitudinal test set.

Figure 6. Gradient visualizations of AI predictions of CKD staging 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, including **a**, Early CKD (corresponding to stage 1 and 2), **b**, Advanced CKD (corresponding to stage 3), and **c**, Severe+ CKD (corresponding to stage 4+). The columns are (1) the original fundus image, (2) a saliency map, and (3) a saliency map overlaying the original image.

Table

Table 1. Basic characteristics of patients in the developmental dataset and external validation cohorts for systemic diseases detection. The numbers of retinal fundus images used for identifying systemic conditions are shown in each cohort. T2DM, Type 2 diabetes mellitus; CKD, chronic kidney disease; eGFR, estimated glomerular filtration rate; DR, diabetic retinopathy; NDR, diabetes mellitus with no DR; BMI，Body mass index.

Cohorts	Developmental Dataset		Internal test set (CC-FII)	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 subjects	30,122	4,307	8,727	8,059	3,081
Male, n (%)	15,325 (50.9%)	2,205 (51.2%)	4,412 (50.6%)	4,441 (55.1%)	1,476 (47.9%)
Age (y), mean (SD)	50.4±14.6	50.6±14.7	50.1±14.6	53.9±13.5	49.0±13.4
BMI (kg/m^2), mean (SD)	24.7±2.4	24.7±2.4	24.7±2.5	24.8±2.6	24.6±2.7
Hypertension, n (%)	9,098 (30.2%)	1,308 (30.4%)	2,601 (29.8%)	2,608 (32.4%)	908 (29.5%)
eGFR (mL/min/1.73 m^2), mean (SD)	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), mean (SD)	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, n (%)	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[a], n (%)	648 (34.0%)	82 (32.7%)	159 (32.9%)	240 (35.5%)	71 (31.1%)
Advanced CKD[b], n (%)	828 (43.4%)	105 (41.8%)	211 (43.6%)	278 (41.1%)	111 (48.7%)
Severe+ CKD[c], n (%)	430 (22.6%)	64 (25.5%)	114 (23.6%)	158 (23.4%)	46 (20.2%)
T2DM, n (%)	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, n (%)	1,414 (16.1%)	228 (17.7%)	392 (16.6%)	425 (15.1%)	141 (11.9%)
T2DM-NDR, n (%)	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 that measurement was available.					

[a]Early CKD is defined as an eGFR of more than 60 mL/min/1.73 m² with albuminuria (corresponding to stage 1 and 2).
[b]Advanced CKD is defined as an eGFR of 30-59 mL/min/1.73 m² (corresponding to stage 3).
[c]Severe+ CKD is defined as an eGFR of less than 30 mL/min/1.73 m² (corresponding to stage 4 and above).

Table 2. Incidence rates of the CKD/T2DM (per 1000 person-year) on the internal longitudinal test set and the external longitudinal test set according to three-strata of the AI models.

Cohorts	Prognostic analysis	Risk group	Low risk	Medium risk	High risk	Overall
Internal longitudinal test set (CC-FII-L), n=3,376	CKD	Number of participants	426	827	432	1685
		Number of events	8	23	49	80
		Incidence rate (per 1,000 person-years, 95% CI)	4.4 (1.9, 8.6)	6.5 (4.1, 9.7)	27.0 (20.0, 35.7)	11.1 (8.8, 13.8)
	T2DM	Number of participants	441	868	469	1778
		Number of events	7	25	57	89
		Incidence rate (per 1,000 person-years, 95% CI)	3.9 (1.6, 8.1)	7.4 (4.8, 11.0)	30.0 (22.7, 38.8)	12.6 (10.2, 15.6)
External longitudinal test set, n=2,112	CKD	Number of participants	359	1016	509	1884
		Number of events	13	27	26	66
		Incidence rate (per 1,000 person-years, 95% CI)	9.4 (5.0, 16.0)	7.3 (4.8, 10.6)	14.5 (9.5, 21.3)	9.6 (7.4, 12.2)
	T2DM	Number of participants	370	1370	1404	3144
		Number of events	4	69	118	191
		Incidence rate (per 1,000 person-years, 95% CI)	2.8 (0.7, 7.1)	14.8 (11.5, 18.7)	25.1 (20.7, 30.0)	17.7 (15.2, 20.3)

Extended Data Tables

Extended Data Table 1. Characteristics of patients in the developmental set and validation sets of two longitudinal cohorts. The numbers of retinal fundus images used for predicting the development of systemic conditions are shown in each cohort. T2DM, Type 2 Diabetes Mellitus; CKD, chronic kidney disease; eGFR, estimated glomerular filtration rate; DR, diabetic retinopathy; NDR, diabetes mellitus with no DR.

Longitudinal Cohorts	Developmental Dataset	Internal longitudinal test set (CC-FII-L)	External longitudinal test set
	Training and Tuning set (CC-FII-L)		
Number of images	16,314	4,224	6,752
Number of subjects	8,157	2,112	3,376
Male, n (%)	3,425 (42.0%)	845 (40.0%)	1,426 (42.2%)
Age (y), mean (SD)	46.2±14.4	46.0±14.5	51.8±13.7
BMI (kg/m^2), mean (SD)	24.1±3.4	24.1±3.5	24.6±3.6
Hypertension, n (%)	2,518 (30.9%)	649 (30.7%)	1,161 (34.4%)
Follow-up time (months), mean (SD)	51.6±15.8	51.6±15.8	51.1±8.5
Participants with known CKD outcomes	6,467	1,685	1,884
Diabetes, n(%)	1,688 (26.1%)	414 (24.6%)	456 (24.2%)
CKD outcome (to Early CKD)	160 (2.5%)	39 (2.3%)	50 (2.7%)
CKD outcome (to Advanced+ CKD)	148 (2.3%)	41 (2.4%)	16 (0.8%)
Participants with known T2DM outcomes	6,807	1,778	3,144
T2DM outcome (to T2DM)	396 (5.8%)	89 (5.0%)	191 (6.1%)

Extended Data Table 2. AI Performance for detection of CKD or T2DM using logistic regression models on internal and external test sets.

Cohorts	Internal test set (CC-FII)	External test set 1	External test set 2: "Point-of-care"
CKD (LR)	0.814 (0.795-0.830)	0.801 (0.785-0.816)	0.784 (0.751-0.813)
T2DM (LR)	0.773 (0.756-0.786)	0.788 (0.767-0.806)	0.774 (0.740-0.802)

Extended Data Table 3. Univariate and multivariate survival analyses of CKD/T2DM conducted using Cox proportional hazards methods (likelihood ratio test).

Prognostic analysis	Covariates	Univariate analysis		Multivariate analysis	
		Hazard ratio	p-value	Hazard ratio	p-value
CKD	**Age**	1.05 (1.05-1.06)	<0.001	1.03 (1.02-1.03)	<0.001
	Sex	0.76 (0.63-0.91)	0.004	0.73 (0.58-0.92)	0.007
	Hypertension	3.53 (2.90-4.30)	<0.001	1.54 (1.23-1.93)	<0.001
	BMI	1.08 (1.05-1.11)	<0.001	1.03 (0.99-1.08)	0.112
	Height	0.97 (0.96-0.98)	<0.001	0.98 (0.97-1.00)	0.059
	Weight	1.01 (1.00-1.01)	0.149	1.00 (0.99-1.02)	0.840
	Diabetes	5.12 (4.17-6.28)	<0.001	2.51 (2.00-3.14)	<0.001
	Fundus (per standard deviation)	3.94 (3.46-4.49)	<0.001	2.10 (1.77-2.50)	<0.001
T2DM	**Age**	1.04 (1.03-1.04)	<0.001	1.02 (1.02-1.03)	<0.001
	Sex	0.65 (0.56-0.76)	<0.001	0.98 (0.79-1.20)	0.823
	Hypertensions	3.24 (2.76-3.79)	<0.001	1.62 (1.36-1.93)	<0.001
	BMI	1.16 (1.14-1.18)	<0.001	1.09 (1.05-1.12)	<0.001
	Height	1.00 (0.99-1.01)	0.807	0.99 (0.98-1.01)	0.425
	Weight	1.03 (1.03-1.04)	<0.001	1.02 (1.01-1.03)	0.002
	Fundus (per standard deviation)	4.35 (3.64-5.19)	<0.001	1.78 (1.37-2.32)	<0.001

Extended Data Table 4. Performance of progression prediction model to CKD or advanced+ CKD event based on the metadata-only model, and the combined model (including fundus images and metadata) on the internal and external test sets. Concordance index (C-index) for right-censored data and 95% confidence intervals (CI) measure the model performance by comparing the progression information (disease labels and progression days) with predicted risk scores. A larger C-index correlates with better progression prediction performance.

Tasks	Progression prediction models	Internal longitudinal test set	External longitudinal test set
CKD	Combined model	0.845 (95% CI: 0.789-0.910)	0.719 (95% CI: 0.627-0.807)
	Metadata model	0.756 (95% CI: 0.699-0.810)	0.651 (95% CI: 0.569-0.730)
Advanced+ CKD	Combined model	0.933 (95% CI: 0.909-0.955)	0.912 (95% CI: 0.823-0.972)
	Metadata model	0.847 (95% CI: 0.804-0.896)	0.832 (95% CI: 0.720-0.924)

Extended Data Table 5. Performance of progression prediction model to T2DM event based on the metadata-only model, and the combined model (including fundus images and metadata) on the internal and external test sets. Concordance index (C-index) for right-censored data and 95% confidence intervals (CI) measure the model performance by comparing the progression information (disease labels and progression days) with predicted risk scores. A larger C-index correlates with better progression prediction performance.

Tasks	Progression prediction models	Internal longitudinal test set	External longitudinal test set
T2DM	Combined model	0.781 (95% CI: 0.743-0.819)	0.765 (95% CI: 0.723-0.799)
	Metadata model	0.774 (95% CI: 0.732-0.819)	0.746 (95% CI: 0.706-0.775)

Extended Data Table 6. Performance of the AI system for CKD detection (including Early CKD, Advanced and Severe+) from normal controls using retinal fundus images. Each row represents metrics based on the corresponding operation point set to perform with high NPV and PPV for CKD screening. CI, confidence interval; PPV, positive predictive value; NPV, negative predictive value.

Operating point based on the tuning set	Cohorts	Sensitivity	Specificity	Reliability of computer-aided decision (CAD)
Positive	Internal test set	43.3% (38.9-49.4)	99.4% (99.1-99.7)	PPV: 92.4% (88.3-95.7)
	External test set 1	34.8% (31.8-38.6)	99.2% (98.9-99.5)	PPV: 88.4% (83.9-92.8)
	External test set 2: "Point-of-care"	29.9% (21.5-34.3)	99.2% (98.5-99.7)	PPV: 89.3% (80.8-95.5)
Negative	Internal test set	99.3% (98.2-100.0)	42.8% (41.0-44.5)	NPV: 99.7% (99.3-100.0)
	External test set 1	99.4% (98.8-99.8)	37.5% (36.1-38.9)	NPV: 99.7% (99.5-99.9)
	External test set 2: "Point-of-care"	99.1% (97.7-100.0)	32.1% (29.0-35.5)	NPV: 99.4% (98.5-100.0)

Extended Data Table 7. Performance of the AI system for T2DM detection using retinal fundus images. Each row represents metrics based on the corresponding operation point set to perform with high NPV and PPV for T2DM screening.
CI, confidence interval; PPV, positive predictive value; NPV, negative predictive value.

Operating point based on the tuning set	Cohorts	Sensitivity	Specificity	Reliability of computer-aided decision (CAD)
Positive	Internal test set	59.1% (54.8-62.0)	97.8% (97.4-98.1)	PPV: 78.7% (75.1-82.1)
	External test set 1	15.9% (12.6-20.7)	99.6% (99.5-99.8)	PPV: 77.9% (71.1-86.2)
	External test set 2: "Point-of-care"	12.1% (8.5-16.1)	99.5% (99.2-99.8)	PPV: 72.7% (59.4-86.1)
Negative	Internal test set	99.3% (98.5-99.8)	41.6% (40.4-43.0)	NPV: 99.8% (99.5-99.9)
	External test set 1	98.8% (97.8-99.7)	31.9% (30.9-33.3)	NPV: 99.7% (99.4-99.9)
	External test set 2: "Point-of-care"	98.5% (96.5-100.0)	44.4% (42.0-46.5)	NPV: 99.6% (99.1-100.0)

Extended Data Table 8. Basic characteristics of patients in the Multi-ethnicity validation cohort for systemic diseases detection. Shown are the numbers of retinal fundus images used for identifying systemic conditions. T2DM, Type 2 diabetes mellitus; CKD, chronic kidney disease; eGFR, estimated glomerular filtration rate; DR, diabetic retinopathy; NDR, diabetes mellitus with no DR; BMI, Body mass index.

Cohort	Multi-ethnicity validation set
Number of images	1,230
Number of subjects	615
Male, n (%)	304 (49.4%)
Age (y), mean (SD)	60.5±12.9
BMI (kg/m^2), mean (SD)	25.3±3.1
Hypertension, n (%)	260 (42.3%)
eGFR (mL/min/1.73 m^2), mean (SD)	86.2±19.0, n=577
Blood glucose (mmol/L), mean (SD)	7.0±2.7, n=586
CKD, n (%)	93 (21.2%), n=439
T2DM, n (%)	343 (55.8%), n=615

Extended Data Figure Legend

Extended Data Figure 1. The flowchart of the AI platform with an ensemble of model instances. We first developed retinal fundus image enhancement models using color normalization and contrast-limited adaptive histogram equalization (CLAHE) techniques. Four types of fundus images after the application of color normalization and CLAHE image enhancements: original image, image after applying the CLAHE transformation only, image after applying the color normalization transformation only, and image after applying both the CLAHE and color normalization transformations. Each image instance separately makes a prediction, and these are combined by averaging the results for producing a robust AI model.

Extended Data Figure 2. **Flow diagram describing the datasets used for our AI system for CKD/T2DM detection and incidence prediction.** Patient inclusion and exclusion criteria were also considered.

Extended Data Figure 3. **Model performance in assessing GFR/CKD staging using retinal fundus images. a-c,** Bland-Altman plot for predicted and actual eGFR after calibrating the model output. AI performance on **a,** the internal test set, **b,** the external test set 1 and **c,** the external test set 2: "point-of-care" study. **d and e,** AI performance in detecting severe+ CKD from other stages of CKD (early and advanced CKD) with the "regression model" and "classification model"**.** The blue curve denoted "classification model" using retinal fundus images. The orange curve denoted "regression model" using thresholds of the predicted GFR from retinal fundus images. **d,** In the internal test. **e,** In the external test set 1. **f**, Correlation analysis of the predicted eGFR of the right eye versus the predicted eGFR of the left eye in normal, early CKD (stages 1 and 2), and CKD. ICC, intraclass correlation coefficient.

Extended Data Figure 4. **Prediction of fasting blood glucose using retinal fundus images. a-c**, Correlation analysis of predicted blood glucose vs actual blood glucose generated using the regression model. **a,** the internal test set, **b,** the external test set 1 and **c,** the external test set 2: "point-of-care" study. **d-f**, Bland-Altman plot for the agreement between the predicted and actual blood glucose levels (mmol/L). Assessing agreement between predicted and estimated blood glucose with mean absolute error (MAE) and intraclass correlation coefficient (ICC)**. d,** the internal test set. **e,** the external test set 1. **f,** the external test set 2: the prospective 'point-of care' pilot study. **g-i,** Bland-Altman plot for predicted and actual blood glucose after calibrating the model output. AI performance on **g,** internal test set, **h,** the external test set 1 and **i,** the external test set 2: the 'point-of-care' study.

Extended Data Figure 5. AI performance at detecting T2DM patients with images with no apparent signs of diabetic retinopathy (NDR) and images with diabetic retinopathy (DR). a and b, Visualizations for detecting T2DM patients by

highlighting the regions the AI model focuses. **a**, T2DM, with "no signs of diabetic retinopathy" (NDR), and **b**, T2DM, with severe DR. The rows are (1) the original fundus image, (2) a saliency map, and (3) a saliency map overlaying the original image. **c,** ROC curves showing performance of binary classification models in the internal test set. Comparison of the AI performance at detecting of T2DM patients with only images with NDR compared to the performance with both images with DR.

Extended Data Figure 6. Performance of the AI model on the external multi-ethnicity validation cohort from Kashi and Macau. a and **b**, ROC curves showing performance of the metadata-only model, the fundus-only model and the combined model on the classification of systematic diseases: **a,** CKD and **b,** T2DM.

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

Extended Data Figure 7. Prediction of the development of CKD and T2DM using the metadata-only model. Kaplan Meier plot illustrating the incidence of CKD/T2DM. The blue, orange, and green lines represent stratified scores for low risk, medium risk, and high risk, respectively, using the upper and lower quartiles in the tuning dataset. **a and b,** progression to CKD on **a,** internal longitudinal test set, **b,** external longitudinal test set. **c and d,** Progression to advanced+ CKD on **c,** internal longitudinal test set, **b,** external longitudinal test set. **e and f,** Progression to T2DM on, **e,** internal longitudinal test set, **f,** external longitudinal test set.

Extended Data Figure 8. The cumulative hazard functions of three stratified risk subgroups (high- medium- and low-risk) using the combined progression prediction model on (a) the internal longitudinal test set and (b) the external longitudinal test set. The solid line is the mean cumulative hazard scores at each time point. The area of the same color represents the 95% confidence interval. Numbers at risk are also included.

Extended Data Figure 9. Prediction of the development of CKD and T2DM using Time-dependent ROC curves. a and b, ROC curves for quantifying AI model performance for the incidence of CKD: **a,** in the internal longitudinal test set for 5 years follow up (case: control=62:470). **b,** in the external longitudinal test set for 4 years follow up (case: control=40:663)**. c and d,** ROC curves for quantifying AI model performance for the incidence of T2DM. **c,** in the internal longitudinal test set

for 5 years follow up (case: control=68:425). **d,** in the external longitudinal test set for 4 years follow up (case: control=96:1,266).

Extended Data Figure 10. **Design Illustration on hand-held smartphone camera attachment.** We used a standard 3D printer to make a customized adaptor that can be fitted and attached to an iPhone X.

Figure 1

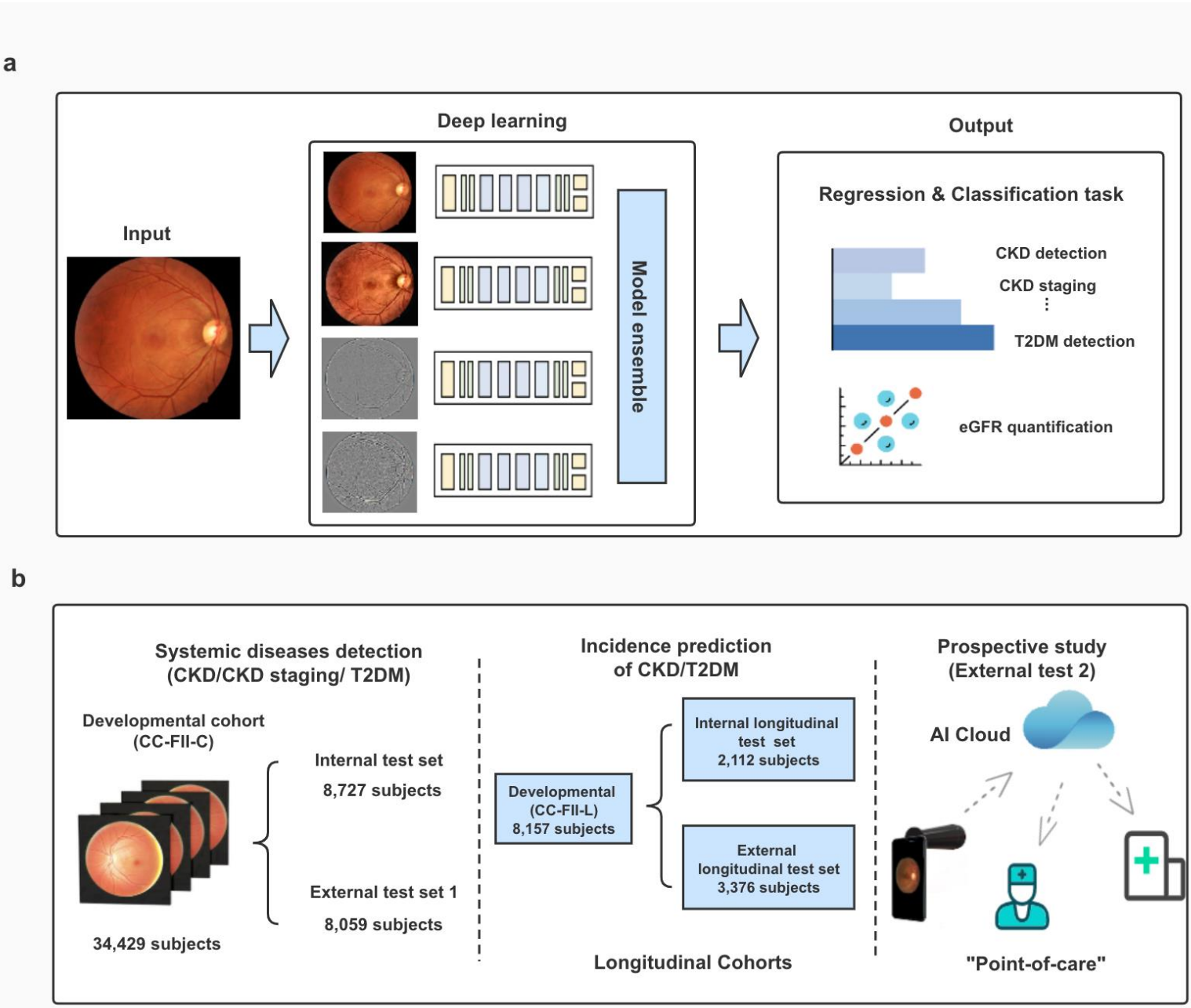

a
Deep learning
Output
Input
Model ensemble
Regression & Classification task
CKD detection
CKD staging
T2DM detection
eGFR quantification
b
Systemic diseases detection
(CKD/CKD staging/ T2DM)
Developmental cohort
(CC-FII-C)
Internal test set
8,727 subjects
External test set 1
8,059 subjects
34,429 subjects
Incidence prediction
of CKD/T2DM
Developmental
(CC-FII-L)
8,157 subjects
Internal longitudinal
test set
2,112 subjects
External
longitudinal test set
3,376 subjects
Longitudinal Cohorts
Prospective study
(External test 2)
AI Cloud
"Point-of-care"

Figure 2

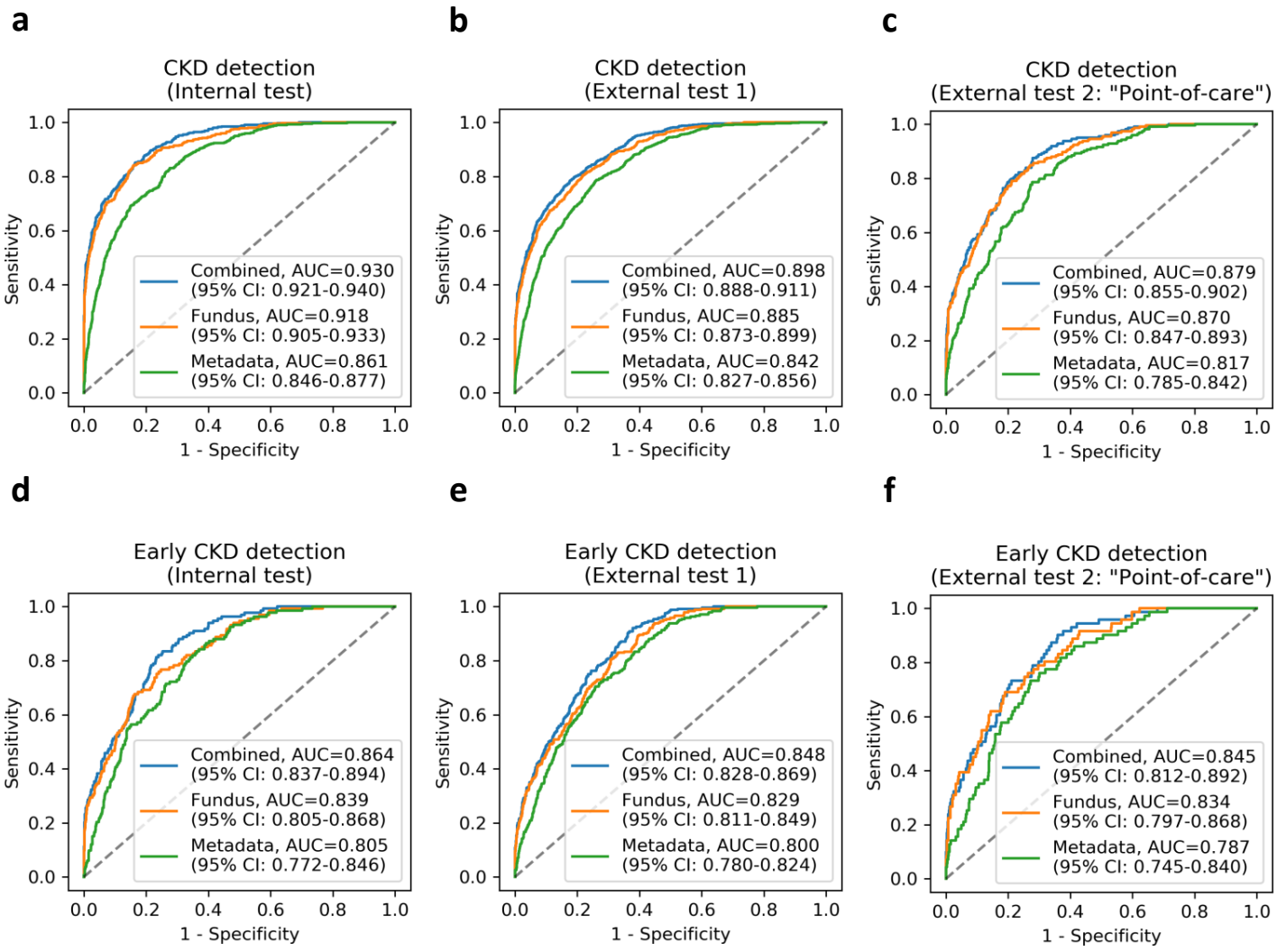

a
CKD detection
(Internal test)
Combined, AUC=0.930
(95% CI: 0.921-0.940)
Fundus, AUC=0.918
(95% CI: 0.905-0.933)
Metadata, AUC=0.861
(95% CI: 0.846-0.877)
b
CKD detection
(External test 1)
Combined, AUC=0.898
(95% CI: 0.888-0.911)
Fundus, AUC=0.885
(95% CI: 0.873-0.899)
Metadata, AUC=0.842
(95% CI: 0.827-0.856)
c
CKD detection
(External test 2: "Point-of-care")
Combined, AUC=0.879
(95% CI: 0.855-0.902)
Fundus, AUC=0.870
(95% CI: 0.847-0.893)
Metadata, AUC=0.817
(95% CI: 0.785-0.842)
d
Early CKD detection
(Internal test)
Combined, AUC=0.864
(95% CI: 0.837-0.894)
Fundus, AUC=0.839
(95% CI: 0.805-0.868)
Metadata, AUC=0.805
(95% CI: 0.772-0.846)
e
Early CKD detection
(External test 1)
Combined, AUC=0.848
(95% CI: 0.828-0.869)
Fundus, AUC=0.829
(95% CI: 0.811-0.849)
Metadata, AUC=0.800
(95% CI: 0.780-0.824)
f
Early CKD detection
(External test 2: "Point-of-care")
Combined, AUC=0.845
(95% CI: 0.812-0.892)
Fundus, AUC=0.834
(95% CI: 0.797-0.868)
Metadata, AUC=0.787
(95% CI: 0.745-0.840)
Sensitivity
1 - Specificity
0.0
0.2
0.4
0.6
0.8
1.0

Figure 3

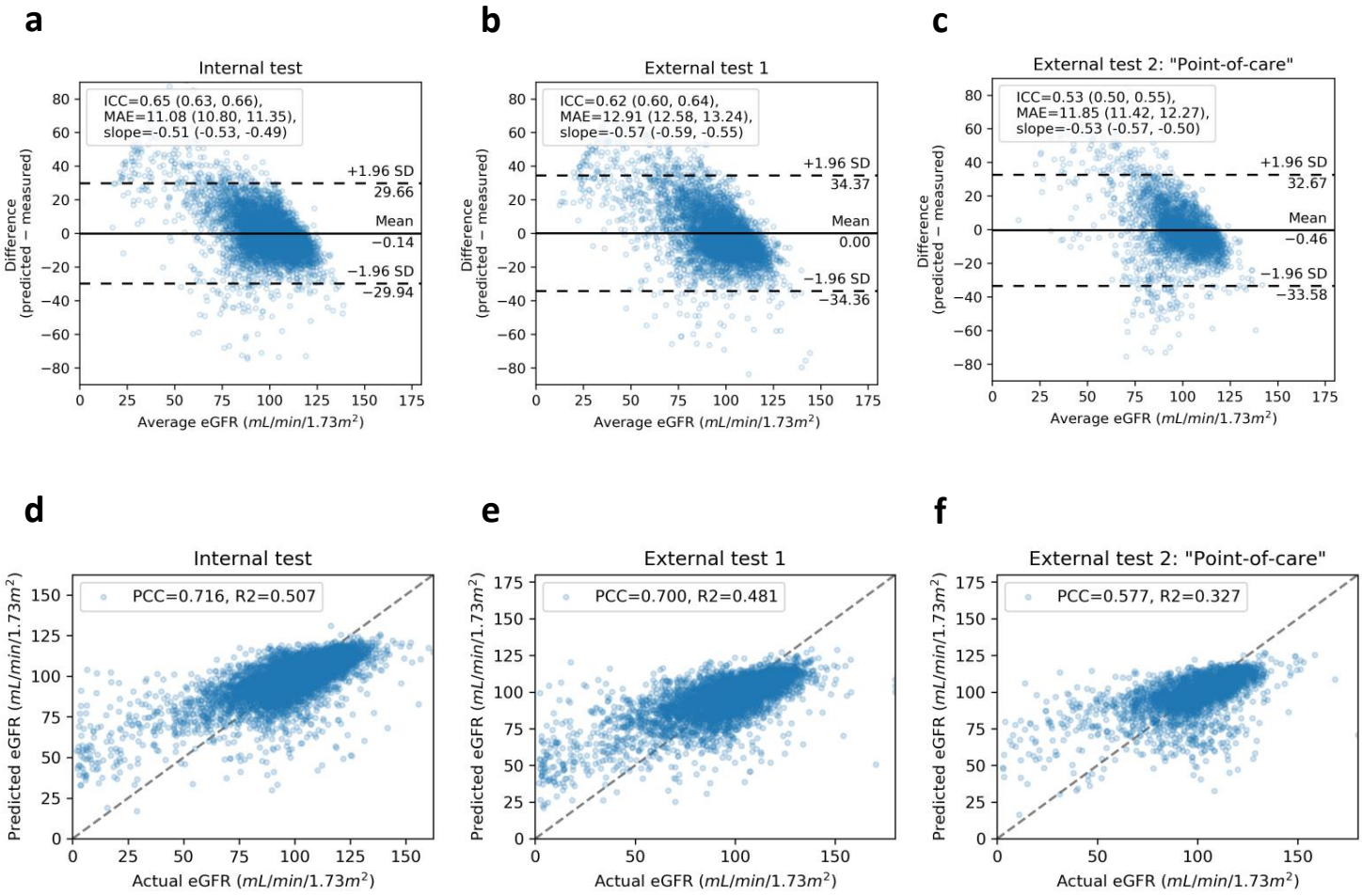
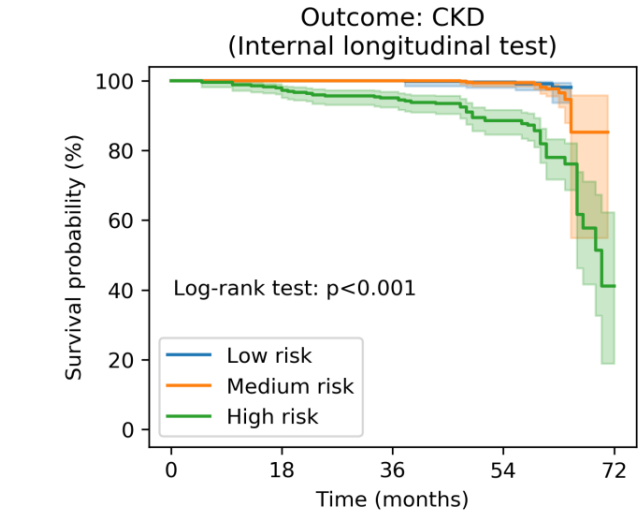

a
Internal test
ICC=0.65 (0.63, 0.66),
MAE=11.08 (10.80, 11.35),
slope=-0.51 (-0.53, -0.49)
+1.96 SD
29.66
Mean
−0.14
−1.96 SD
−29.94
Difference
(predicted − measured)
Average eGFR (mL/min/1.73m²)
b
External test 1
ICC=0.62 (0.60, 0.64),
MAE=12.91 (12.58, 13.24),
slope=-0.57 (-0.59, -0.55)
+1.96 SD
34.37
Mean
0.00
−1.96 SD
−34.36
Difference
(predicted − measured)
Average eGFR (mL/min/1.73m²)
c
External test 2: "Point-of-care"
ICC=0.53 (0.50, 0.55),
MAE=11.85 (11.42, 12.27),
slope=-0.53 (-0.57, -0.50)
+1.96 SD
32.67
Mean
−0.46
−1.96 SD
−33.58
Difference
(predicted − measured)
Average eGFR (mL/min/1.73m²)
d
Internal test
PCC=0.716, R2=0.507
Predicted eGFR (mL/min/1.73m²)
Actual eGFR (mL/min/1.73m²)
e
External test 1
PCC=0.700, R2=0.481
Predicted eGFR (mL/min/1.73m²)
Actual eGFR (mL/min/1.73m²)
f
External test 2: "Point-of-care"
PCC=0.577, R2=0.327
Predicted eGFR (mL/min/1.73m²)
Actual eGFR (mL/min/1.73m²)

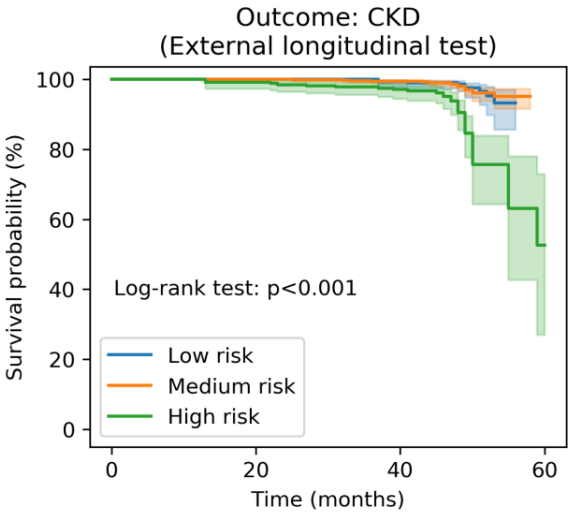
Figure 4

a



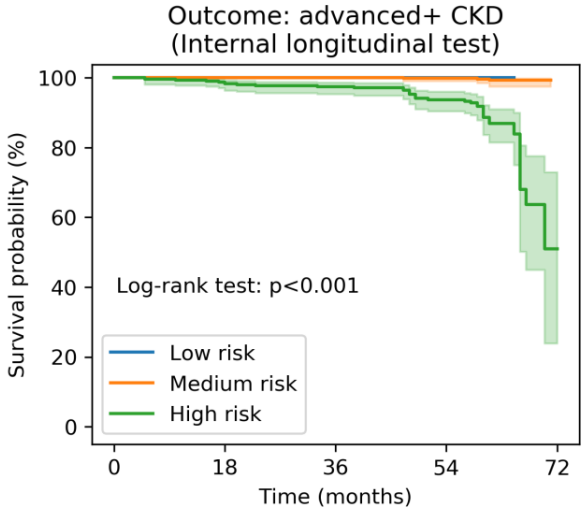
	T=0	T=18	T=36	T=54	T=72
Low risk	459	459	459	270	0
Medium risk	815	808	802	457	0
High risk	409	370	327	202	2

b



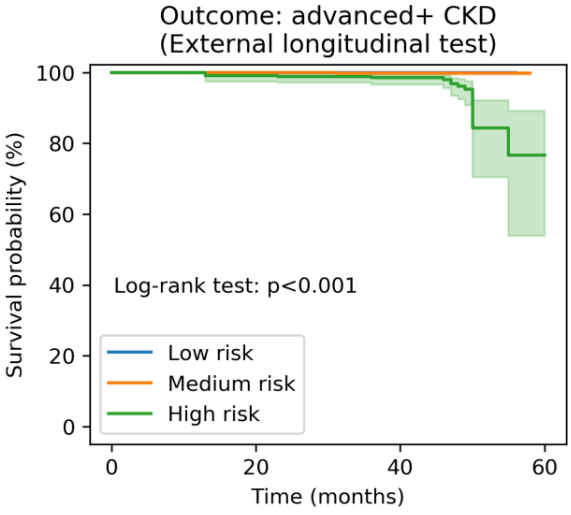
	T=0	T=20	T=40	T=60
Low risk	396	396	341	0
Medium risk	1145	1137	905	0
High risk	342	322	231	5

c



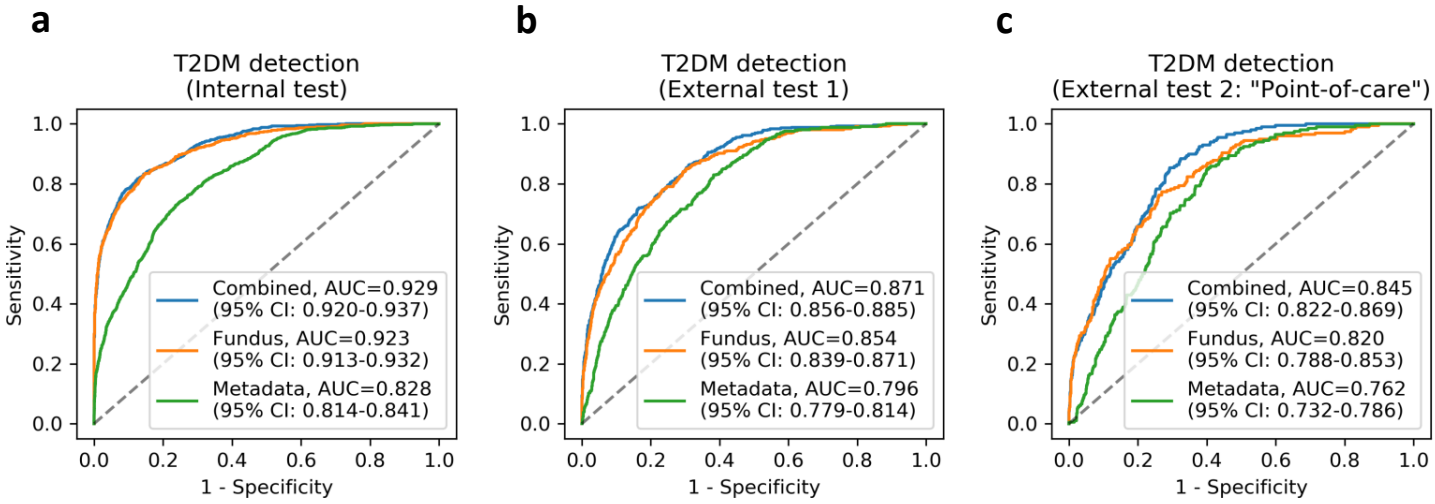
	T=0	T=18	T=36	T=54	T=72
Low risk	442	442	442	239	0
Medium risk	835	821	813	476	0
High risk	407	375	334	214	2

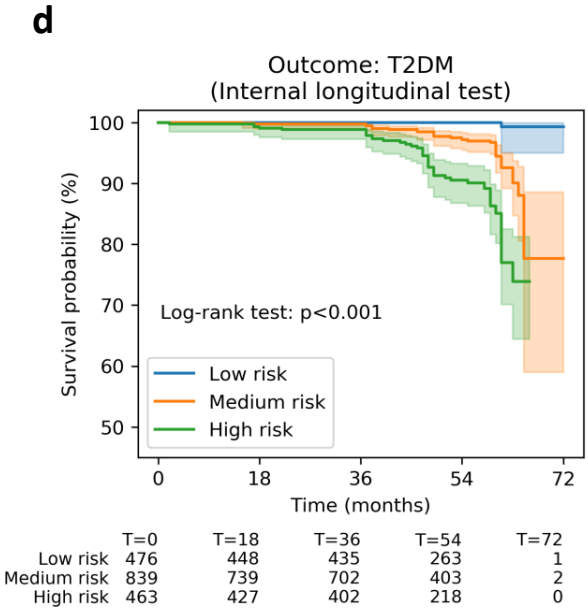
d



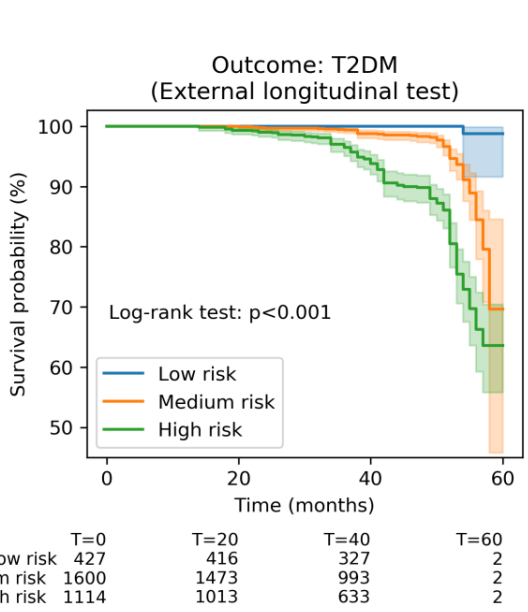
	T=0	T=20	T=40	T=60
Low risk	377	377	324	0
Medium risk	1134	1125	912	0
High risk	373	354	242	5

Figure 5





	T=0	T=18	T=36	T=54	T=72
Low risk	476	448	435	263	1
Medium risk	839	739	702	403	2
High risk	463	427	402	218	0



	T=0	T=20	T=40	T=60
Low risk	427	416	327	2
Medium risk	1600	1473	993	2
High risk	1114	1013	633	2

Figure 6