Supplementary information

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

In the format provided by the authors and unedited

Table of contents

Supplementary Methods

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 (Supplementary Fig.12). 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 or a commercially available Volk Optical iNvew hand-held fundoscope (https://www.volk.com/products/inview-for-iphone-6-6s) 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. Images with clarity and inclusion of key anatomical landmarks such as the optic nerve, macula and posterior arcade retinal vessels, were selected for the study.
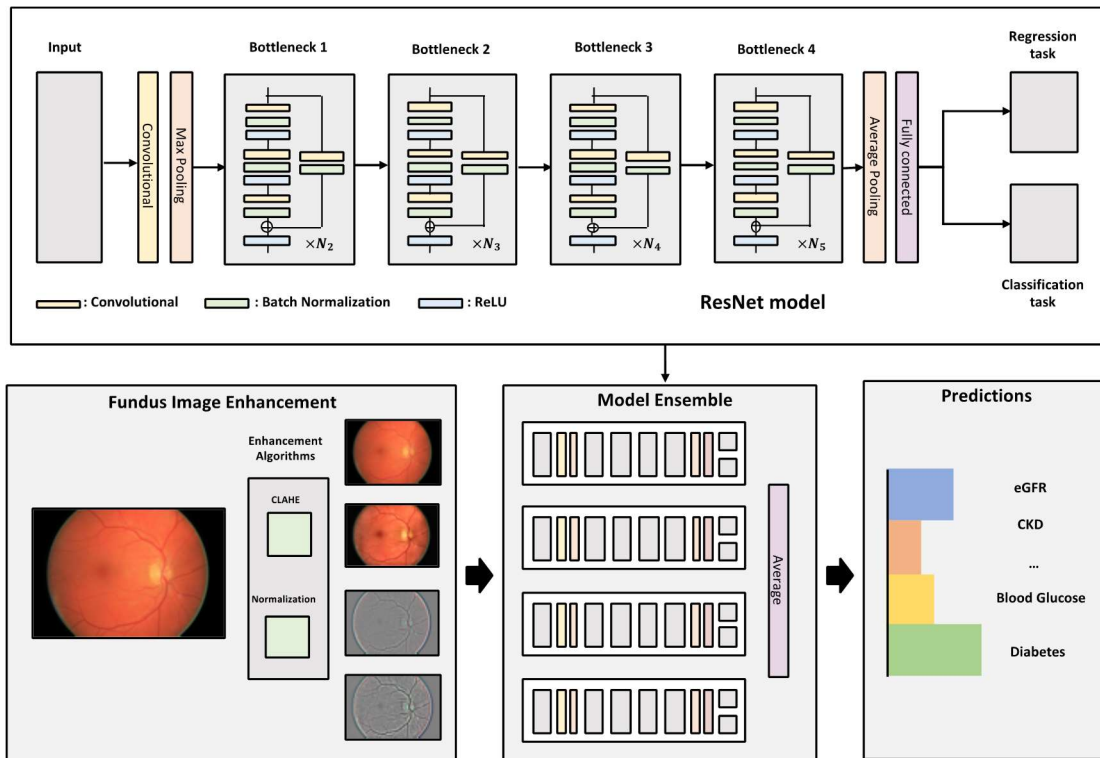
Supplementary Figure 1 | The flowchart of the AI platform with an ensemble of model instances.

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 to produce a robust AI model.
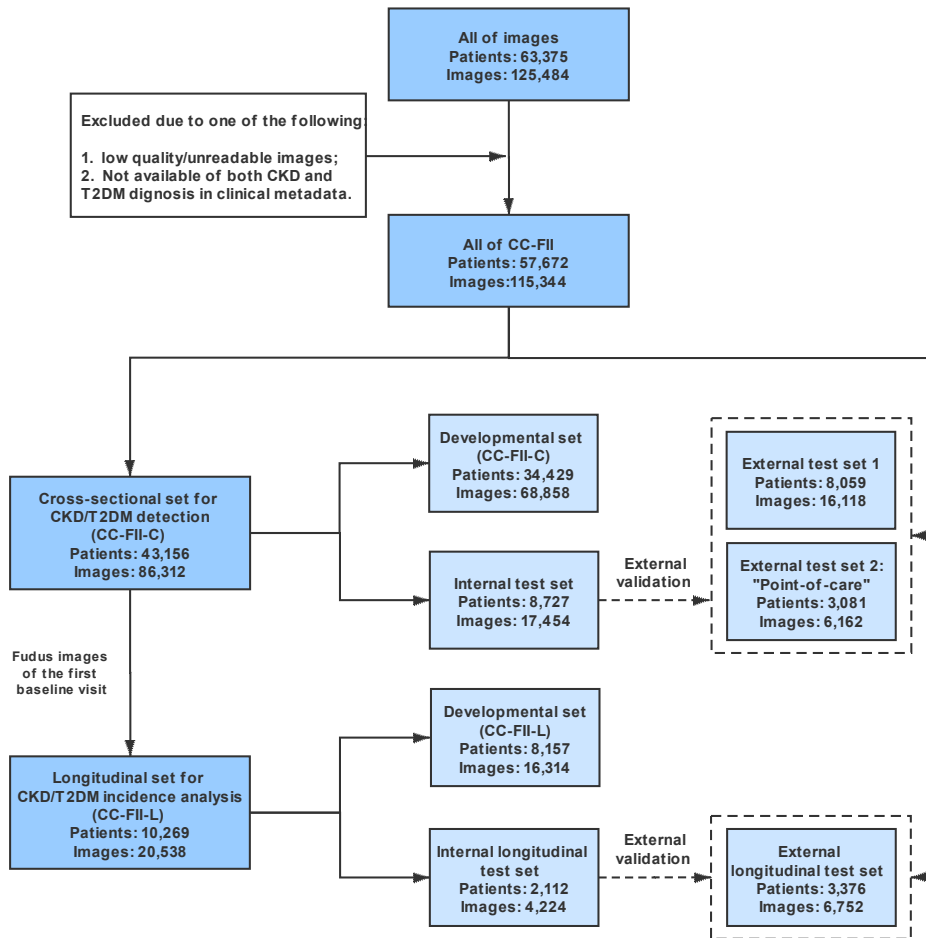
Supplementary Figure 2 | Flow diagram describing the datasets used for our AI system for CKD/T2DM detection and incidence prediction.
Patient inclusion and exclusion criteria were also considered.
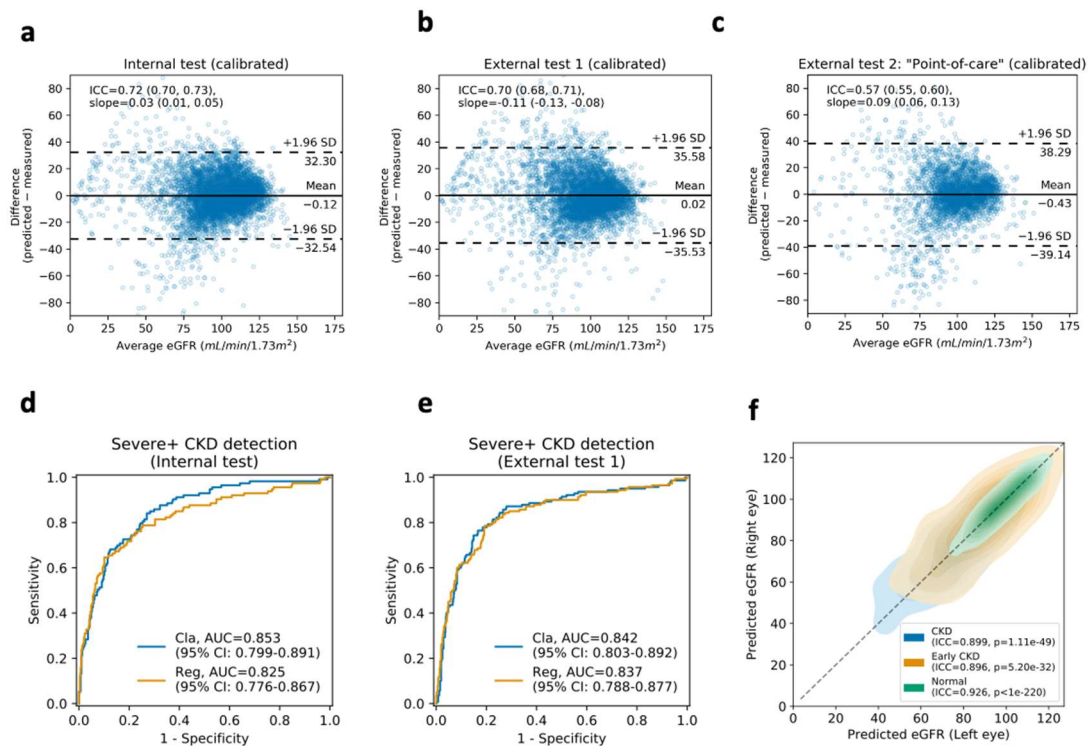
Supplementary Figure 3 | Model performance in assessing GFR/CKD staging using retinal fundus images.

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" in **d,** the internal test. **e,** the external test set 1. 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. **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; CI, confidence intervals.
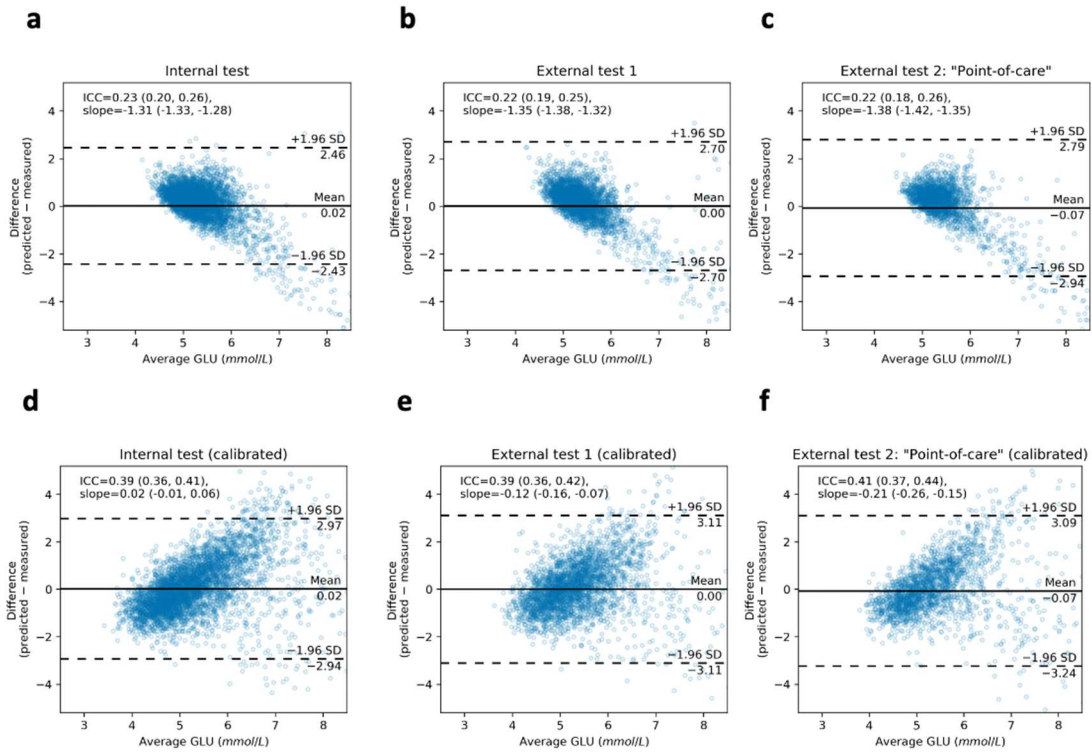
Supplementary Figure 4 | Prediction of fasting blood glucose using retinal fundus images.

a-c, Bland-Altman plot for the agreement between the predicted and actual blood glucose levels (mmol/L). The performance of AI system on **a,** the internal test set. **b,** the external test set 1. **c,** the external test set 2: the prospective 'point-of care' pilot study. **d-f,** Bland-Altman plot for predicted and actual blood glucose after calibrating the model output. The performance of AI system on **d,** internal test set, **e,** the external test set 1 and **f,** the external test set 2: the 'point-of-care' study. ICC, intraclass correlation coefficient.
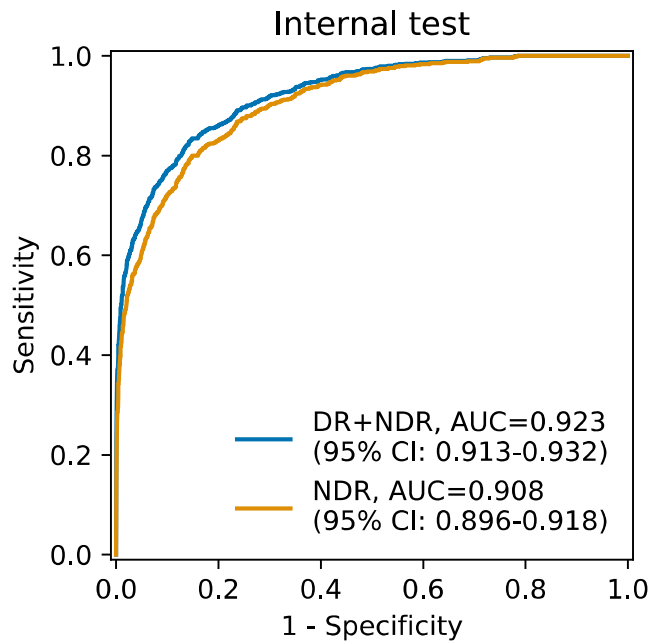
Supplementary Figure 5 | Comparison of the AI's performance at detecting T2DM patients with images with no apparent signs of diabetic retinopathy (NDR) and images with diabetic retinopathy (DR).
ROC curves showing performance of binary classification models in the internal test set. The orange line represents T2DM patients with only images with NDR. The blue line represents T2DM patients with both images with DR and NDR.

a

CKD detection
(Multi-ethnicity validation set)

Combined, AUC=0.876
(95% CI: 0.845-0.906)
Fundus, AUC=0.842
(95% CI: 0.803-0.882)
Metadata, AUC=0.817
(95% CI: 0.777-0.858)

b

T2DM detection
(Multi-ethnicity validation set)

Combined, AUC=0.859
(95% CI: 0.827-0.886)
Fundus, AUC=0.800
(95% CI: 0.760-0.832)
Metadata, AUC=0.798
(95% CI: 0.761-0.828)
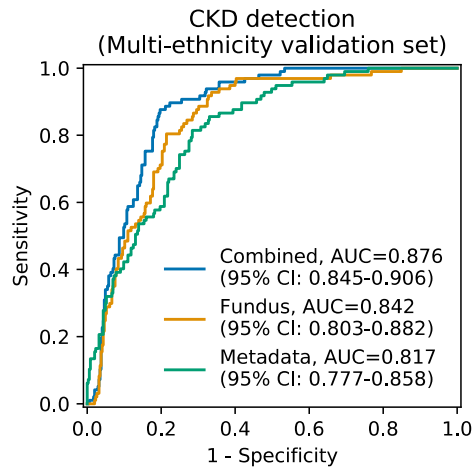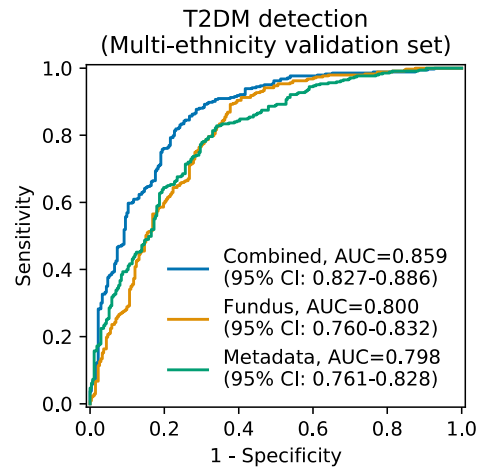
Supplementary Figure 6 | Performance of the AI system on the external multi-ethnicity validation cohort from Kashi and Macau.
a and **b**, ROC curves showing performance of the metadata-only model, the fundus-only model and the combined model on the classification of systematic diseases: **a,** CKD and **b,** T2DM.
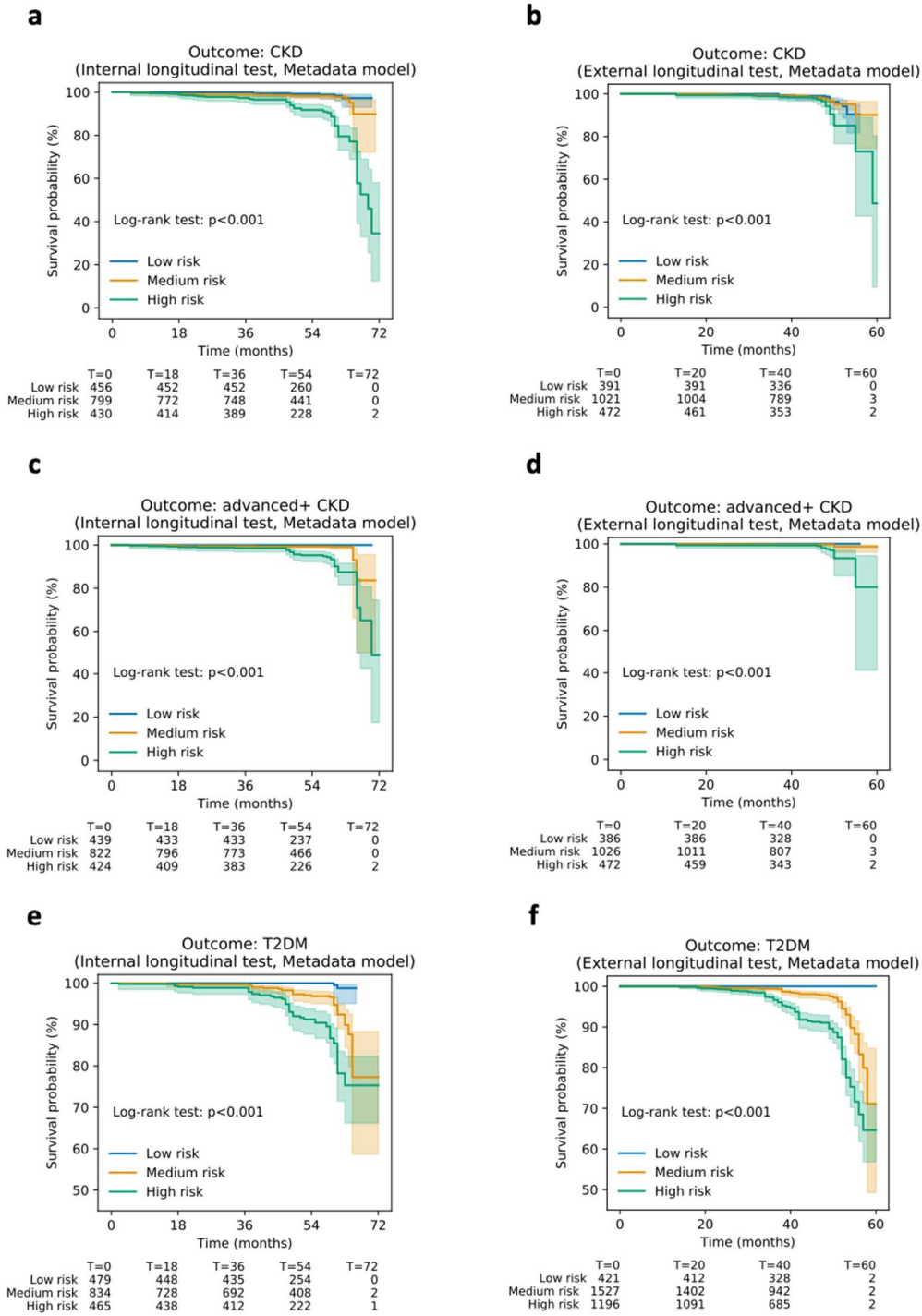
Supplementary Figure 7 | Kaplan Meier plot illustrating the incidence of CKD/T2DM using the metadata-only model.
The blue, orange, and green lines represent stratified scores for low risk, medium risk, and high risk, respectively. The area of the same color represents the 95% confidence interval. The tables below represent the number of patients at risk at a particular time point stratified by the risk levels. **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, **d,** external longitudinal test set. **e** and **f,** Progression to T2DM on, **e,** internal longitudinal test set, **f,** external longitudinal test set. P-value is computed using a one-sided log-rank test between all groups.

a

Outcome: CKD
(Internal longitudinal test)

	T=0	T=18	T=36	T=54	T=72
Low risk	460	460	460	270	0
Medium risk	815	808	802	457	0
High risk	410	370	327	202	2

b

Outcome: CKD
(External longitudinal test)

	T=0	T=20	T=40	T=60
Low risk	397	397	342	0
Medium risk	1145	1137	905	0
High risk	342	322	231	5

c

Outcome: advanced+ CKD
(Internal longitudinal test)

	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	408	375	334	214	2

d

Outcome: advanced+ CKD
(External longitudinal test)

	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

e

Outcome: T2DM
(Internal longitudinal test)

	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

f

Outcome: T2DM
(External longitudinal test)

	T=0	T=20	T=40	T=60
Low risk	430	419	329	2
Medium risk	1600	1473	993	2
High risk	1114	1013	633	2

Supplementary Figure 8 | The cumulative hazard functions of three stratified risk subgroups (high- medium- and low-risk) using the combined progression prediction model.
The solid line is the mean cumulative hazard scores at each time point. The area of the same color represents the 95% confidence interval. The tables below represent the number of patients at risk at a particular time point stratified by the risk levels. **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, **d,** external longitudinal test set. **e** and **f,** Progression to T2DM on, **e,** internal longitudinal test set, **f,** external longitudinal test set.
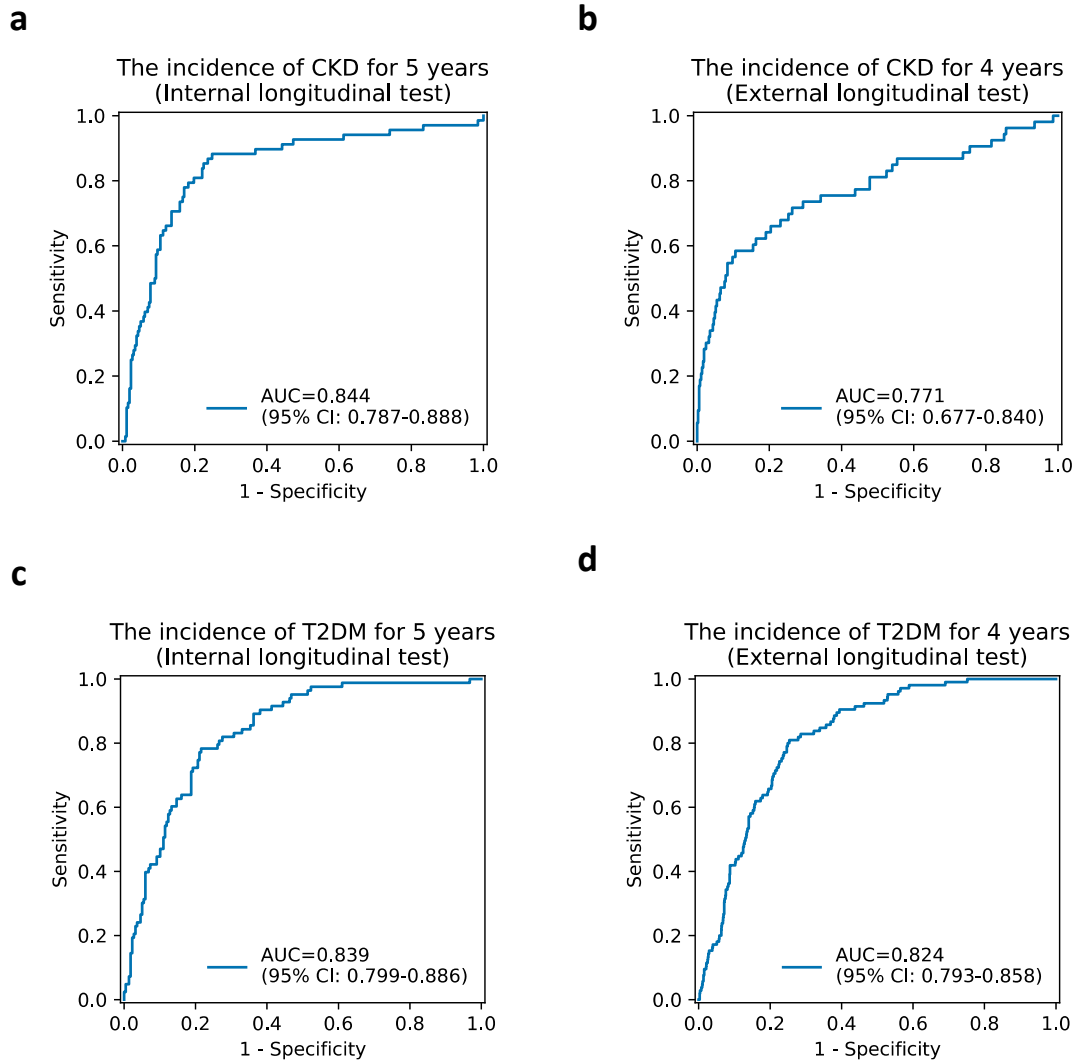
Supplementary Figure 9 | Prediction of the development of CKD and T2DM using time-dependent ROC curves.
a and **b,** ROC curves for quantifying AI model performance for the incidence of CKD in **a,** the internal longitudinal test set for 5 years follow up (case: control=62:470). **b,** the external 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 in **c,** the internal longitudinal test set for 5 years follow up (case: control=68:425). **d,** the external longitudinal test set for 4 years follow up (case: control=96:1,266).

a CKD detection

Dx'ed+unDx'ed, AUC=0.917	(95% CI: 0.902-0.926)
unDx'ed, AUC=0.905	(95% CI: 0.881-0.924)

b T2DM detection

Dx'ed+unDx'ed, AUC=0.911	(95% CI: 0.900-0.920)
unDx'ed, AUC=0.915	(95% CI: 0.904-0.927)
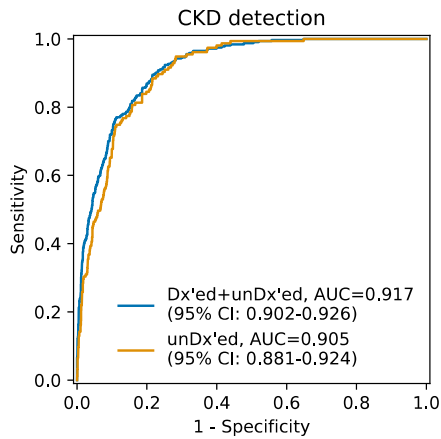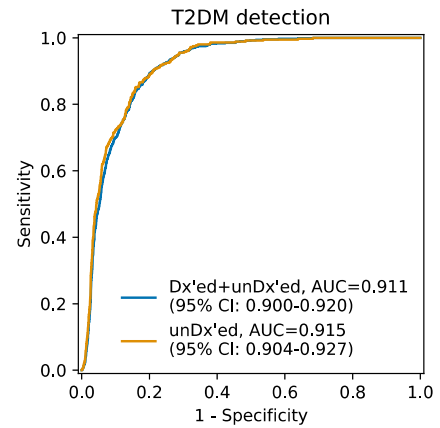
Supplementary Figure 10 | Performance of the AI models on identifying documented CKD/T2DM of the AI models in test sets.
(i) The blue line represents a mixed cohort (all) including patients with previously diagnosed (Dx'ed) and previously undiagnosed (unDx'ed) disease; (ii) The orange line represents a cohort excluding previously diagnosed (Dx'ed) disease. **a,** ROC curves of CKD detection performance in (i) a mixed cohort (including Dx'ed CKD) (case:control ratio = 314:2,685); (ii) a cohort without previously diagnosed (unDx'ed) CKD (case:control ratio = 155:2,685). **b,** ROC curves of T2DM detection performance in (i) a mixed cohort (including Dx'ed T2DM) (case:control ratio = 672:2,685); (ii) a cohort without previously diagnosed (unDx'ed) T2DM (case:control ratio = 358:6,366).
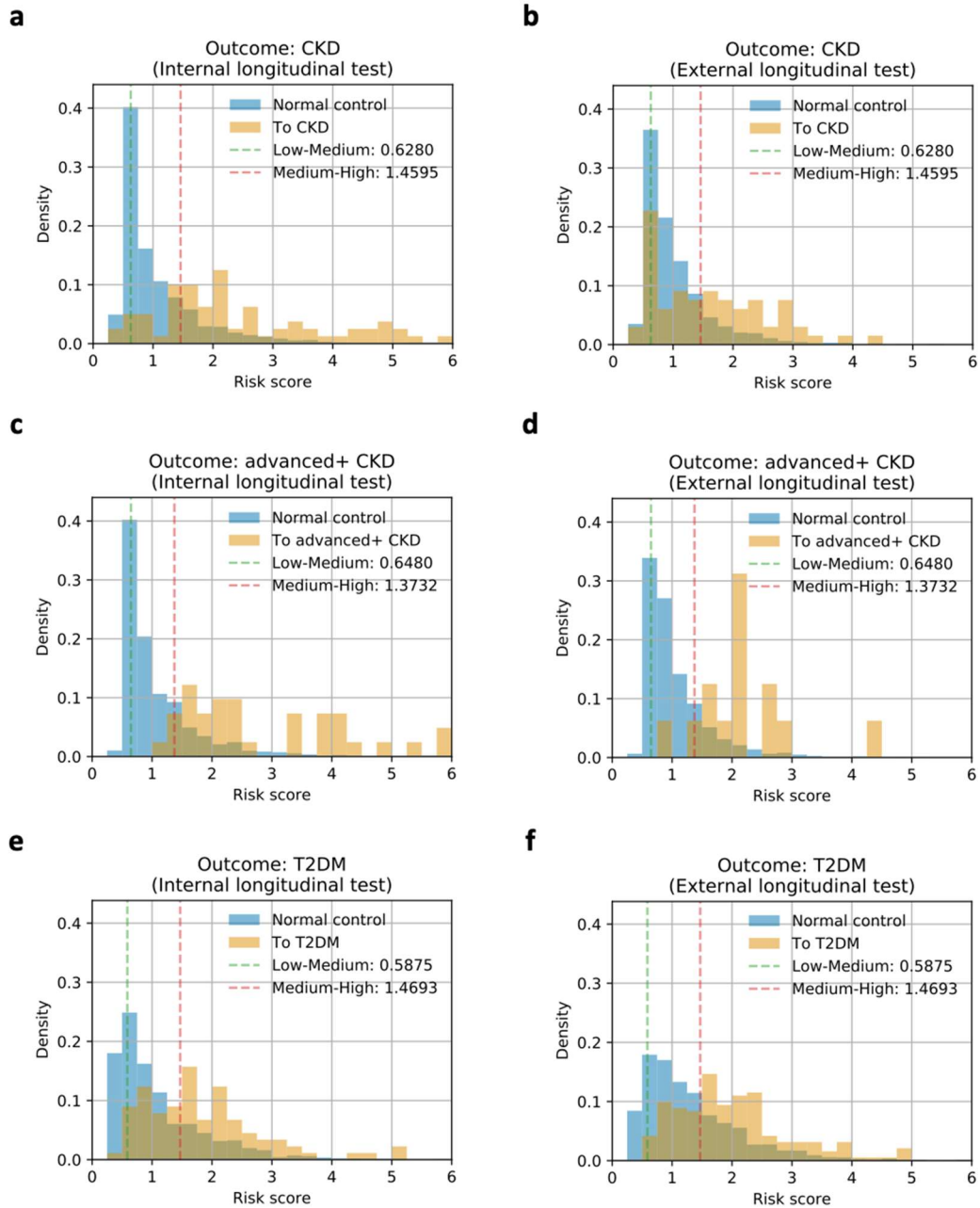
Supplementary Figure 11 | The distribution of risk scores of the prognostic models across all datasets.

The green dot line represents the the low-medium threshold of risk score. The red dot line represents the medium-high threshold of risk score. **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.

Supplementary Figure 12 | 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.

Supplementary 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 Training and Tuning set (CC-FII-L)	Internal longitudinal test set (CC-FII-L)	External longitudinal test set
Number of images	16,314	4,224	6,752
Number of participants	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%)

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

Cohorts	Internal test set	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)

Supplementary 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 (1.04-1.06)	<0.001	1.02 (1.01-1.03)	<0.001
	Sex	0.75 (0.62-0.91)	0.003	0.76 (0.60-0.96)	0.019
	Hypertension	3.14 (2.58-3.81)	<0.001	1.36 (1.09-1.71)	0.008
	BMI	1.07 (1.04-1.10)	<0.001	1.03 (0.99-1.07)	0.149
	Height	0.97 (0.96-0.98)	<0.001	0.99 (0.97-1.00)	0.128
	Weight	1.01 (1.00-1.01)	0.152	1.00 (0.99-1.01)	0.801
	Diabetes	5.05 (4.12-6.19)	<0.001	2.54 (2.03-3.19)	<0.001
	Fundus (per standard deviation)	4.06 (3.55-4.63)	<0.001	2.25 (1.89-2.69)	<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	1.01 (0.82-1.24)	0.947
	Hypertension	3.24 (2.76-3.79)	<0.001	1.35 (1.11-1.63)	0.002
	BMI	1.16 (1.14-1.18)	<0.001	1.08 (1.04-1.11)	<0.001
	Height	1.00 (0.99-1.01)	0.807	0.99 (0.98-1.01)	0.353
	Weight	1.03 (1.03-1.04)	<0.001	1.02 (1.00-1.03)	0.005
	Fundus (per standard deviation)	4.39 (3.67-5.24)	<0.001	1.76 (1.36-2.28)	<0.001

Supplementary Table 4 | Incidence rates of the Advanced+ CKD (per 1000 person-year) on the internal longitudinal test set and the external longitudinal test set according to three-strata of the AI models.

Risk group	Number of participants	Number of events	Incidence rate (95% CI)	Univariate analysis		Multivariate analysis	
				HR (95% CI)	P-value	HR (95% CI)	P-value
Internal longitudinal test set (CC-FII-L)							
Low risk	442	0	0.0 (0.0, 0.0)	NA	NA	NA	NA
Medium risk	835	3	0.8 (0.2, 2.4)	Reference	NA	Reference	NA
High risk	408	38	23.4 (16.6, 32.2)	8.17 (4.35, 15.34)	<0.001	3.22 (1.48, 7.01)	0.003
Overall[a]	1685	41	5.7 (4.1, 7.7)	4.36 (3.04, 6.23)	<0.001	1.81 (1.10, 2.98)	0.019
External longitudinal test set							
Low risk	377	0	0.0 (0.0, 0.0)	NA	NA	NA	NA
Medium risk	1134	1	0.2 (0.0, 1.3)	Reference	NA	Reference	NA
High risk	373	15	11.9 (6.7, 19.6)	6.74 (3.14, 14.48)	<0.001	3.54 (1.46, 8.56)	0.005
Overall[a]	1884	16	2.3 (1.3, 3.8)	5.27 (2.89, 9.63)	<0.001	3.38 (1.53, 7.51)	0.003
P-values were computed using the likelihood ratio test. HR, hazard ratio; CI, confidence interval.							
[a]A continuous variable was used (predicted z-score).							

Supplementary Table 5 | 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% 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. CI, confidence interval.

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)

Supplementary Table 6 | 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% 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. CI, confidence interval.

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)

Supplementary Table 7 | Numbers at risk of the Kaplan Meier plots illustrating the incidence of CKD/T2DM stratified by three risk subgroups (high- medium- and low-risk).
T2DM, Type 2 diabetes mellitus; CKD, chronic kidney disease. T is the follow-up time (months).

Outcome	Risk group	Internal longitudinal test set					External longitudinal test set			
		T=0	T=18	T=36	T=54	T=72	T=0	T=20	T=40	T=60
CKD	Low risk	460	460	460	270	0	397	397	342	0
	Medium risk	815	808	802	457	0	1145	1137	905	0
	High risk	410	370	327	202	2	342	322	231	5
Advanced+ CKD	Low risk	442	442	442	239	0	377	377	324	0
	Medium risk	835	821	813	476	0	1134	1125	912	0
	High risk	408	375	334	214	2	373	354	242	5
T2DM	Low risk	476	448	435	263	1	430	419	329	2
	Medium risk	839	739	702	403	2	1600	1473	993	2
	High risk	463	427	402	218	0	1114	1013	633	2

Supplementary Table 8 | Performance of the AI system for CKD detection 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. 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)

Supplementary Table 9 | 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. 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)

Supplementary Table 10 | Basic characteristics of patients in the Multi-ethnicity validation cohort for systemic diseases detection.
Shown are the number 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 participants	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
Blood glucose (mmol/L), mean (SD)	7.0±2.7
CKD, n (%)	93 (21.2%), n=439
T2DM, n (%)	343 (55.8%), n=615