A deep learning algorithm to detect chronic kidney disease from retinal photographs in community-based populations

Charumathi Sabanayagam, Dejiang Xu, Daniel SW Ting, Simon Nusinovici, Riswana Banu, Haslina Hamzah, Cynthia Lim, Yih-Chung Tham, Carol Y Cheung, E Shyong Tai, Ya Xing Wang, Jost B Jonas, Ching-Yu Cheng, Mong Li Lee, Wynne Hsu, Tien Y Wong

Summary

Background Screening for chronic kidney disease is a challenge in community and primary care settings, even in high-income countries. We developed an artificial intelligence deep learning algorithm (DLA) to detect chronic kidney disease from retinal images, which could add to existing chronic kidney disease screening strategies.

Methods We used data from three population-based, multiethnic, cross-sectional studies in Singapore and China. The Singapore Epidemiology of Eye Diseases study (SEED, patients aged ≥40 years) was used to develop (5188 patients) and validate (1297 patients) the DLA. External testing was done on two independent datasets: the Singapore Prospective Study Program (SP2, 3735 patients aged ≥25 years) and the Beijing Eye Study (BES, 1538 patients aged ≥40 years). Chronic kidney disease was defined as estimated glomerular filtration rate less than 60 mL/min per 1·73m². Three models were trained: 1) image DLA; 2) risk factors (RF) including age, sex, ethnicity, diabetes, and hypertension; and 3) hybrid DLA combining image and RF. Model performances were evaluated using the area under the receiver operating characteristic curve (AUC).

Findings In the SEED validation dataset, the AUC was 0·911 for image DLA (95% CI 0·886–0·936), 0·916 for RF (0·891–0·941), and 0·938 for hybrid DLA (0·917–0·959). Corresponding estimates in the SP2 testing dataset were 0·733 for image DLA (95% CI 0·696–0·770), 0·829 for RF (0·797–0·861), and 0·810 for hybrid DLA (0·776–0·844); and in the BES testing dataset estimates were 0·835 for image DLA (0·767–0·903), 0·887 for RF (0·828–0·946), and 0·858 for hybrid DLA (0·794–0·922). AUC estimates were similar in subgroups of people with diabetes (image DLA 0·889 [95% CI 0·850–0·928], RF 0·899 [0·862–0·936], hybrid 0·925 [0·893–0·957]) and hypertension (image DLA 0·889 [95% CI 0·860–0·918], RF 0·889 [0·860–0·918], hybrid 0·918 [0·893–0·943]).

Interpretation A retinal image DLA shows good performance for estimating chronic kidney disease, underlying the feasibility of using retinal photography as an adjunctive or opportunistic screening tool for chronic kidney disease in community populations.

Funding National Medical Research Council, Singapore.

Copyright © 2020 The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY 4.0 license.

Introduction Chronic kidney disease is a major global public health problem. Because earlier detection of chronic kidney disease allows appropriate interventions, regular screening is recommended for the general population,1,2 and in high-risk populations (eg, patients with diabetes or hypertension, and specific ethnic groups).3 However, screening of chronic kidney disease depends on measurement of the estimated glomerular filtration rate (eGFR, calculated from serum creatinine), or urine tests for protein or albumin. Because serum or urine samples must be obtained, adherence to screening is low, even in high-income countries and in at-risk populations. A study in Australia showed that nearly 50% of patients with diabetes attending general practice had not been screened for chronic kidney disease in the previous 18 months.4 Although a urine sample is less invasive and easier to obtain, albuminuria is highly variable, with intra-individual variation of up to 50%.4

The kidney and eye share similar structural, developmental, physiological (renin–angiotensin–aldosterone hormonal cascade), and pathogenic pathways (inflammation, oxidative stress, endothelial dysfunction and microangiopathy).5 Patients with clinically visible retinal microvascular signs (eg, retinopathy, arteriolar narrowing, venular dilatation) are more likely to have chronic kidney disease,6,7 which suggests that the retina could provide screening information to complement existing methods. Digital retinal photography is non-invasive and commonly used in community and primary care settings for eye screening, particularly for diabetic retinopathy.

Artificial intelligence (AI) deep learning algorithms (DLAs) applied to imaging data have shown superior performance to manual interpretation of images in diagnosing conditions such as diabetic retinopathy6,8 and skin cancer.9 DLAs applied to kidney ultrasonography might also detect chronic kidney disease.10 In this study, we developed and validated a DLA for predicting chronic
kidney disease from retinal images and compared this with two DLA models, one using classic clinical risk factor (RF) data and another using both retinal and RF data.

**Methods**

We did a conventional development, validation, and external testing study on three DLAs (retinal images only, RF only, and combined retinal and RF hybrid) using retinal images and clinical data collected from three population-based studies.14–18 We developed and internally validated the DLAs using data from the Singapore Epidemiology of Eye Diseases (SEED) study,14–16 and externally tested the DLAs on two independent datasets: the Singapore Prospective Study Program (SP2)17 and the Beijing Eye Study (BES).18

**Datasets**

For DLA development, data and images were obtained from participants who attended the baseline (2004–11) and 6-year follow-up (2011–17) visits of SEED, a population-based study of Chinese, Malay, and Indian participants aged 40 years or older at baseline. Detailed methods of this study have been published (appendix p 1).14–16 Two datasets were used for external testing of the DLA: 1) SP217 (2004–07, 4098 patients), a population-based study of Chinese, Malay, and Indian participants aged 24 years or older at baseline and 2) BES18 (2001, 4439 patients), a population-based study of Chinese participants aged 40 years or older at baseline (2001, 4439 patients).

The study was done in accordance with the tenets of the Declaration of Helsinki and ethics approval was obtained from the SingHealth Institutional Review Board. Written informed consent was provided by all participants.

**Definition of chronic kidney disease cases and controls**

We defined chronic kidney disease as an eGFR of less than 60 mL/min per 1·73 m² (corresponding to stage 3 and above) and controls as those with eGFR 60 mL/min per 1·73 m² or higher (corresponding to chronic kidney disease stage 0–2). eGFR was calculated from serum creatinine using the Chronic Kidney Disease Epidemiology Collaboration equation.19 In the development SEED dataset, we assessed presence and absence of chronic kidney disease in both visits. Cases were those who had chronic kidney disease in either visit but without overlaps. If a participant had chronic kidney disease in one of the visits, images from the corresponding visit were used, and for participants with chronic kidney disease in both visits, only images from baseline were used for development. Controls were defined as those who had no chronic kidney disease in both visits (6191 participants). For all controls, images from baseline were used. Combining both visits, 2027 participants (1217 from SEED-1 and 810 from SEED-2) had chronic kidney disease. From this pool of cases and controls (appendix p 2), after excluding those with missing or poor quality images, missing data on suggesting that the retina could provide clues for chronic kidney disease.

**Evidence before this study**

Chronic kidney disease is a major cause of preventable morbidity and mortality, and screening is recommended for at-risk populations such as in people with diabetes or hypertension. Screening of chronic kidney disease relies on measuring serum concentration of creatinine or testing urine for protein, which might not be practical in all settings. Other screening approaches have been explored, including using clinical risk factor model-based prediction scores before evaluating with serum creatinine. The Screening for Occult Renal Disease model using several demographic and clinical variables showed modest performance with an area under the curve (AUC) of 0·88 in internal validation model and 0·71 in external validation model. Artificial intelligence-based deep learning algorithms (DLAs) using imaging data have been shown to have superior performance in diagnosing conditions like diabetic retinopathy and skin cancer. Retinal photography is non-invasive and is increasingly commonly used in the community and primary care setting to screen for eye diseases, particularly in patients with diabetes. The retina shares a close biological relationship with the kidney, and retinal microvascular abnormalities like retinopathy and other vascular features have been shown to be associated with chronic kidney disease, even in people without diabetes.

**Added value of this study**

We developed and validated a DLA to screen for chronic kidney disease non-invasively from retinal images, using data from a large population-based study, and externally validated the DLA in two independent datasets in Singapore and China. The DLA showed good performance with AUC of 0·911 in internal validation and 0·733 and 0·835 in external tests sets. The performance of the image DLA in subgroups of participants with diabetes (AUC 0·889) and hypertension (0·889) were similar to that of the whole group.

**Implications of all the available evidence**

This is the first study to link the retina and the kidney with an Artificial intelligence-based DLA and shows the potential of retinal images as a tool to detect and screen chronic kidney disease with good accuracy in the community. DLAs have the potential to be integrated into retinal cameras to serve as a complementary community-based or primary care-based model for chronic kidney disease screening, which is traditionally reliant only on serum creatinine and estimated glomerular filtration rate.
classic RF, and overlapping cases, data or images from 6485 participants (1218 cases, 5267 controls) were used for developing the algorithm.

Definition of chronic kidney disease in SP2 and BES was similar to that of the SEED study. After excluding participants with missing or poor quality retinal images, and missing data on serum creatinine, data from 3735 participants in SP2 (240 cases, 3495 controls) and 1538 participants in BES (53 cases, 1485 controls) were used for external testing.

Retinal imaging
In all three studies, two-field (optic disc centred and macula centred) retinal photography was taken from each eye after pupil dilation according to the Early Treatment for Diabetic Retinopathy Study protocol using a 45° non-mydriatic digital retinal camera (Canon CR-DGi with a 10D/20D SLR backing, Canon, Japan in SEED and SP221 and CR6-4SNM, Canon, Japan in BES).22

RF
We used five classic RF (age, sex, ethnicity, diabetes and hypertension) as predictors for the RF model. Age, sex, and ethnicity were self-reported and information on diabetes and hypertension were obtained from self-report, physical examination, and laboratory examination in all datasets (appendix p 1).

Algorithm development
We used 12 970 retinal images from 6485 SEED participants to train and test the DLAs. The DLA takes as inputs two standardised macula-centred images (1 image per eye per participant) with resolution of 512 × 512. We also developed DLAs separately using macula-centred images, optic disc-centred images, and both (combination). The DLA using combined images (both macula-centred and optic disc-centred) performed slightly better (area under the receiver operating characteristic curve [AUC] for image-only in SEED 0.9017) than using either one separately (AUC 0.9015 for macula-centred and 0.8926 for optic disc-centred [data not shown]). However, because not all participants in the external test sets had both views, but all had macula-centred views, we opted to use macula-centred images only for the final DLA.

The output for the DLA was a binary classifier with the predicted chronic kidney disease status. The deep learning model is based on cCondenseNet24 with 5 blocks, as shown in figure 1. During the training process, the parameters of the network are initially set to random values. For each image, the prediction given by the neural network is compared with its ground truth label, and parameters are updated to reduce the prediction error. We used the 5-fold cross-validation approach to evaluate the performance of the model by dividing the whole SEED dataset into 5 parts preserving the proportion of chronic kidney disease cases and controls as the original dataset. Each time, one part was held out for validation and the remaining 4 parts (80%) were used for developing the model. The validation set had no overlap with the training set. The performance of the trained DLA was evaluated on the validation set to classify chronic kidney disease status by averaging the AUCs across the 5 test samples.

Figure 1: Convolutional neural network architecture for chronic kidney disease from retinal images

![Convolutional neural network architecture](image)

Input was a concatenation of two normalised cropped retinal images, one macula-centred image per eye, and the output was a fully connected layer with the predicted chronic kidney disease status. The architecture comprises of 5 dense blocks alternated with transition layers to down-sample the features. Each dense block is a series of cAdd units,20 which is a new propagation mechanism of deepening and widening the neural architecture with two types of convolutions (1 × 1 and 3 × 3). Every transition layer is a 1 × 1 convolution with pooling. cAdd/channel-wise-addition.

Table 1: Baseline characteristics of participants

| Dataset          | SEED training set (n=5188) | SEED validation set (n=1297) | SP2 validation cohort (n=3735) | BES validation cohort (n=1538) |
|------------------|-----------------------------|--------------------------------|-------------------------------|-------------------------------|
| Number of images| 10 376                      | 2594                          | 7470                          | 3076                          |
| Chronic kidney disease |                              |                                |                               |                               |
| Stage 3         | 881 (SD 90.5)               | 219 (SD 89.8)                 | 225 (SD 93.8)                 | 52 (SD 98.1)                  |
| Stage 4         | 66 (SD 6.7)                 | 15 (SD 6.1)                   | 12 (SD 5.0)                   | 1 (SD 1.9)                    |
| Stage 5         | 27 (SD 2.8)                 | 10 (SD 4.1)                   | 3 (SD 1.3)                    | 0                             |
| Age (years)     | 58.4 (SD 9.9)               | 58.4 (SD 9.9)                 | 49.9 (SD 11.5)                | 64.3 (SD 9.6)                 |
| Ethnicity       |                              |                                |                               |                               |
| Male            | 2742 (52.9%)                | 666 (51.3%)                   | 1793 (48.0%)                  | 616 (41.4%)                   |
| Female          | 2446 (47.1%)                | 631 (48.7%)                   | 1942 (52.0%)                  | 902 (58.6%)                   |
| Diabetes        |                              |                                |                               |                               |
| Chinese         | 1486 (28.6%)                | 415 (32.0%)                   | 367 (9.8)                     | 259 (16.8%)                   |
| Malay, Indian   | 824 (62.5%)                 | 824 (63.5%)                   | 1518 (40.6)                   | 785 (51.0)                    |
| Hypertension    |                              |                                |                               |                               |
| Chinese         | 82.6 (SD 22.1)              | 81.8 (SD 22.0)                | 87.7 (SD 17.7)                | 89.6 (SD 14.3)                |
| Malay, Indian   | 82.4 (63.5%)                | 82.4 (63.5%)                  | 1518 (40.6)                   | 785 (51.0)                    |
| eGFR (mL/min per 1.73 m²) | 82.6 (SD 22.1) | 81.8 (SD 22.0) | 87.7 (SD 17.7) | 89.6 (SD 14.3) |

Data are n (%) or mean (SD). SEED=Singapore Epidemiology of Eye Diseases. SP2=Singapore Prospective Study Program. BES=Beijing Eye Study. eGFR=estimated glomerular filtration rate.
Statistical analysis

We developed 3 different DLA models: 1) retinal image DLA 2) RF DLA (age, sex, ethnicity, diabetes, hypertension status) and 3) retinal image and RF (hybrid) DLA. The performance of the 3 models were evaluated in the internal validation set (SEED) and the external test sets using AUC. We measured sensitivity and specificity at an optimal threshold balancing the two measures and calculated positive predictive value and negative predictive value. To identify the most important regions in an image with respect to the classification (ie, how the DLA arrived at prediction), we generated heatmaps.

We performed sensitivity analyses to test the robustness of the algorithm. First, since misclassification of chronic kidney disease would be common among those with eGFR near to normal range (ie, 55–60 mL/min per 1·73 m²), we tested the DLA using an alternate definition of chronic kidney disease (eGFR <45 mL/min per 1·73 m² corresponding to stage 3B and higher). Secondly, since the DLA was developed using a general population cohort (SEED), we tested the DLA in high-risk subgroups, including those with diabetes and hypertension, separately. For this analysis, we used only the SEED validation set since few patients in SP2 and BES had diabetes and chronic kidney disease or hypertension and chronic kidney disease. We also did two supplementary analyses. First we repeated our imaging DLAs using pre-set high sensitivity (95%) and pre-set high specificity (95%). Secondly, in addition to AUC, we estimated area under the precision-recall curve (AUPRC), an alternative metric suggested for class imbalance data. Similar to AUC, AUPRC values range from 0 to 1, but while a value of 0·5 indicates a random guess for AUC, AUPRC is dependent on prevalence, and tends to be 0 when the prevalence is low.

Role of the funding source

The funders had no role in the study design, data collection, analysis, preparation of the manuscript, and decision to publish. The corresponding author had full access to all data and final responsibility for the decision to submit for publication.

Results

Participant characteristics are shown in table 1. Characteristics of the participants in the SEED training and validation sets were similar, except for a higher prevalence of diabetes in the validation set (415 [32·0%] of 1297 participants) versus training set (1486 [28·6%] of 5188 participants). Participants in BES were older (64·3 years [SD 9·6]) than those in SEED (58·4 years [SD 9·9]) and SP2 (49·9 years [SD 11·5]). Prevalence of diabetes and hypertension were higher and mean eGFRs were lower in SEED participants than in SP2 and BES participants. Most cases in all 3 datasets (90–98%) were
stage 3 (eGFR 60–30 mL/min per 1·73 m²). 4·1% of cases in SEED, 1·3% in SP2, and no cases in BES were stage 5 (eGFR <15 mL/min per 1·73 m²).

AUC of the DLAs for image-only, RF, and the hybrid model are shown in figure 2. In SEED, AUC was 0·911 for image-only (95% CI 0·886–0·936), 0·916 for RF (0·891–0·941), and 0·938 for the hybrid model (0·917–0·959, p=0·001). AUC estimates were 0·733 in SP2 (0·696–0·770) and 0·835 in BES (0·767–0·903) for image-only, which improved to 0·829 in SP2 (0·797–0·861) and 0·887 in BES (0·828–0·946) for RF. In the hybrid model, AUC was higher than image-only DLA in both SP2 (0·810 vs 0·733; p=0·0005) and BES (0·858 vs 0·835; p=0·0002). AUCs in the hybrid model were lower than that in the RF model in SP2 (0·810 vs 0·829) and BES (0·858 vs 0·835).

Performance metrics comparing the 3 models in internal validation and the external test sets are shown in table 2. In SEED, the DLA had 83% sensitivity and 83% specificity for image. Sensitivity and specificity were similar for RF and the hybrid model. In SP2, both sensitivity and specificity were 70% for image; 73% sensitivity, 80% specificity for RF; and 74% sensitivity, 75% specificity for the hybrid model. Sensitivity and specificity estimates in BES were 75% each for image; 79% sensitivity, 82% specificity for RF; and 79% each for the hybrid model. For image-only DLA, the negative predictive value was more than 95% in all three datasets across all 3 models. However, positive predictive value for image was very low in SP2 (14%) and BES (9%), compared with the SEED validation set (54%). Positive predictive value estimate for the hybrid model was 16% in SP2, 57% in SEED, and 11% in BES.

In sensitivity analyses, using an alternative definition of chronic kidney disease (defined as eGFR <45 ml/min), the performance of the DLA improved significantly for all models in all datasets (table 3). For image-only DLA, AUC estimates improved from 0·911 to 0·933 in SEED validation, 0·733 to 0·827 in SP2, and 0·835 to 0·904 in BES. Performance of the DLA in subgroups of diabetes and hypertension in SEED validation were similar to the main model (AUC for diabetes was 0·889 for image, 0·899 for RF, and 0·925 for hybrid; AUC for hypertension was 0·889 for image, 0·889 for RF, and 0·918 for hybrid).

In pre-set high sensitivity (95%) analyses, the negative predictive value for image was high in all three datasets (98% in SEED and SP2 and approximately 100% in BES). In pre-set high specificity analyses, positive predictive value increased from 54% to 76% in SEED, 14% to 22% in SP2, and 9% to 20% in BES. Estimated AUPRC for image-only for SEED validation was good (0·793, 95% CI 0·758–0·828) but, as expected, it was lower for SP2 (0·167, 0·147–0·187), and BES (0·140, 0·105–0·175; data not shown).

In heatmaps (figure 3), changes related to dilatation of venules, rarefaction of vessels, and retinopathy changes were noted in DLA positive cases (figure 3C–E), suggesting that the signals for the prediction could be distributed throughout the image.

Table 3: Performance of the deep learning algorithm in sensitivity analyses

| Case: control | Image-only (95% CI) | Risk factor only (95% CI) | Hybrid (95% CI) |
|---------------|---------------------|--------------------------|----------------|
| Chronic kidney disease defined by eGFR <45 mL/min |
| SEED validation | 90:1053 | 0·933 (0·897–0·969) | 0·938 (0·903–0·973) | 0·960 (0·932–0·988) |
| SP2 | 53:3495 | 0·827 (0·758–0·896) | 0·889 (0·831–0·947) | 0·867 (0·805–0·929) |
| BES | 14:1485 | 0·904 (0·798–1·00) | 1·000 (1·00–1·00) | 0·982 (0·933–1·033) |
| Subgroup with diabetes |
| SEED validation | 120:285 | 0·889 (0·850–0·928) | 0·899 (0·862–0·936) | 0·925 (0·893–0·957) |
| Subgroup with hypertension |
| SEED validation | 227:597 | 0·889 (0·860–0·918) | 0·889 (0·860–0·918) | 0·918 (0·893–0·943) |

BES=Beijing Eye Study. eGFR=estimated glomerular filtration rate. SEED=Singapore Epidemiology of Eye Diseases. SP2=Singapore Prospective Study Program.
Discussion

We developed and validated a retinal image-based DLA for detecting chronic kidney disease, with the aim to determine if this approach is feasible for an adjunctive or opportunistic screening in community settings. Our results indicate that chronic kidney disease can be accurately detected from retinal images without knowledge of specific retinal signs (eg, retinopathy, figure 3). The image-only DLA and clinical RF models both achieved high AUCs in SEED internal validation, and modest to good AUCs in external test sets of SP2 and BES. However, the addition of RF to retinal image (to form the hybrid DLA) only modestly improved the prediction of chronic kidney disease, and only in the SEED validation, suggesting that if retinal photography alone were used for screening, even simple risk factor information need not be collected from the population. The performance of image DLA in subgroups of patients with diabetes and hypertension was similar to that in the whole group. Thus, for chronic kidney disease detection, a retinal image-only DLA is similar to information from a classic RF model, and supports the potential of using retinal photography to detect chronic kidney disease in specific settings (eg, community-based, primary care, or high-risk groups). This offers a novel opportunity to complement chronic kidney disease screening in different clinical and public health settings.

Chronic kidney disease has a long asymptomatic phase and is amenable to screening, so preventive strategies could reduce the health-care burden substantially. Moreover, awareness of the disease among those with chronic kidney disease is low, at 10%.21 However, guidelines for chronic kidney disease screening are not consistent across professional bodies. The American Society of Nephrology recommends routine screening (including those at risk, including people with diabetes, hypertension, cardiovascular disease, or family history of chronic kidney disease, and ethnic groups at high risk).21 The at-risk approach misses cases; a study in Norway24 found high-risk chronic kidney disease screening strategies that targeted diabetes or hypertension missed more than half of chronic kidney disease cases. A systematic review25 evaluating the cost-effectiveness of primary screening for chronic kidney disease suggested that screening was cost-effective, particularly in people with diabetes or hypertension, with incremental cost-effectiveness ratios ranging from US$100 253–109 912 per quality-adjusted life-year (QALY) for eGFR-based screening and $14063–160 018 per QALY for proteinuria-based screening. Studies are needed to identify the benefits and risks of screening, screening measures, and target groups for screening of asymptomatic people.25 Few risk prediction models have been developed for screening chronic kidney disease.26–28 Most have been based on traditional risk factors, except for one model10 based on deep learning, using ultrasound images of kidney to predict chronic kidney disease. The Screening for Occult Renal Disease model developed using data from the US National Health and Nutrition Examination Survey (1999–02), and externally validated in the Atherosclerosis Risk in Communities study population achieved an AUC of 0·88 and 0·71 in internal and external validation models.26 The model was developed based on several clinical variables associated with chronic kidney disease like diabetes, hypertension, anaemia, history of cardiovascular disease, peripheral vascular disease, congestive cardiac failure, and proteinuria. A point-of-care screening model developed in India using age, sex, waist circumference, body-mass index, and urine dipstick achieved a C-statistic of 0·76 in internal validation and 0·74 and 0·70 in external test sets.27 The deep learning model based on kidney ultrasonography developed in Taiwan achieved an AUC of 0·904 with 92·1% specificity but only 60–7% sensitivity, and lacked external validation.28 In our study, prediction based on retinal image alone performed very well in internal validation and had reasonable performance in the external test sets.

While the negative predictive value of image-only DLA was more than 95% in all datasets, the positive predictive value was high in SEED validation, but low in SP2 and BES, because positive predictive value is influenced by the prevalence of chronic kidney disease in the population. Prevalence of chronic kidney disease in SEED was high (baseline prevalence was 12%), and we combined cases from two visits to balance cases and controls. Thus, proportion of chronic kidney disease cases in the SEED validation set was 23%. Prevalence of chronic kidney disease in SP2 and BES was low, resulting in low positive predictive value. To increase positive predictive value in populations with low prevalence of chronic kidney disease, we recommend the image-only DLA be applied to at-risk groups, such as people with diabetes or hypertension. The current image-only DLA showed good performance in SEED because of the high prevalence of hypertension, diabetes, and the age of participants. Positive predictive value could also be improved by choosing a high-specificity operating threshold. For the current image-only DLA, we chose a threshold that had optimal sensitivity and specificity (both >80%). If we choose a 95% specificity cutoff, positive predictive value improved in all 3 datasets. Depending upon the needs of the health-care system, different operating thresholds can be chosen. Since the retinal image-based DLA has similar performance to that of classic RF models, we also suggest applying this tool as part of a two-stage or sequential screening as used in diabetes screening where fasting blood glucose is used to screen first followed by oral glucose tolerance. Being non-invasive, image-only DLA
can be applied first and those who screen positive are recalled for further testing with serum creatinine, reducing the number of false positives.

To overcome the limitations of class imbalance, alternative metrics such as kappa statistic and AUPRC have been suggested over accuracy and AUC. Kappa, a traditional measure of agreement between two raters (inter-rater agreement) has widely been used as a classification performance metric in machine learning literature. In unbalanced situations, distribution of misclassification has been shown to affect the value of kappa to the extent that worse classification results can occur despite high values of kappa, and Delgado and colleagues discouraged using kappa as a performance measure comparing classifiers. In low-prevalence settings, AUPRC has been suggested to overcome the optimism of ROC, thus reporting AUPRC as a supplement to AUC. AUPRC has been suggested to overcome the optimism of comparing classifiers. In low-prevalence settings, AUPRC has been suggested to overcome the optimism of ROC, thus reporting AUPRC as a supplement to AUC. AUPRC focuses on positive predictive value (precision) and sensitivity (recall) but completely ignores true negatives, which is the dominant group in low-prevalence disease. In our study, AUC and PRC performance were quite good in SEED validation, but AUC performance was modest, with a low positive predictive value and a low PRC (<0.2) in both SP2 and BES.

A crucial element of any AI-based prediction tool is its clinical relevance. In clinical practice, when chronic kidney disease screening is recommended by physicians, blood or urine is taken on the same or another day and the results are sent several hours later to the physician. The patient is scheduled to return another day, when the physician reviews the result with the patient. In addition, fear of needles is a common problem that affects around 10% of individuals and leads to avoidance of medical care. Thus, despite the low cost and easy availability of serum creatinine testing, compliance to chronic kidney disease screening is low, even in high-income countries and in people with diabetes.

The retinal changes that the AI or DLA is picking up probably represent the cumulative effects of multiple shared pathways leading to chronic kidney disease, including microvascular damage resulting from age, diabetes, hypertension, and inflammation. The microvascular changes representing vessel dilatation, rarefaction, and retinopathy observed in heatmaps of DLA positive cases were consistent with similar quantitative findings reported previously. Since the fundus image-based DLA has similar performance to that of classic RF model, it can be used as an alternative screening model. The use of fundus-image-based DLA has similar performance to that of classic RF model, it can be used as an alternative screening model. However, the algorithm that detects characteristic microvascular changes and uses these features to predict chronic kidney disease could potentially solve this issue. We are also developing a feature-based deep learning retinal vessel algorithm for predicting chronic kidney disease.

In conclusion, our study shows the potential of a DLA using retinal images to identify chronic kidney disease in community populations. Since access to digital retinal photography is increasing at the community and primary care level, a retinal image-based DLA has the potential to be adopted for first stage chronic kidney disease screening before confirmatory tests.

**Contributors**

CS and TYW had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. CS, MLL, WH and TYW conceived and designed the study. DX, SN and CS analysed the data. CS and RB drafted the manuscript. DX, DSWT, SN, HH, CL, Y-CT, CYC, EST, YXW, JBJ, C-YC, MLL, WH, and TYW critically revised the manuscript for important intellectual content. Final version of the paper has been seen and approved by all the authors.

**Declaration of interests**

TYW reports consultancy and payments for lectures from Allergan, Bayer, Boehringer Ingelheim, Denentina, Merck, Novartis, Oxurion, and Chronic Kidney Disease Screening could be combined with diabetic retinopathy screening in diabetic patients, potentially increasing the chronic kidney disease screening at population level.

The major strengths of this study are the development and validation of an algorithm based on retinal image only, which is simple and easy to obtain at primary care or community level. We were also able to have two independent population-based cohorts with similar imaging and chronic kidney disease protocol to validate our algorithm.

Our study has several limitations. First, we did not have information on albuminuria for all participants, thus we could not evaluate an albuminuria-based prediction DLA. Second, chronic kidney disease was defined based on a single measurement of eGFR and this could have resulted in misclassification of chronic kidney disease cases. Third, although heatmaps indicated vessel changes and abnormal lesions characteristic of retinopathy, we do not know what features were used by the DLA to identify chronic kidney disease. A multistep algorithm that detects characteristic microvascular changes and uses these features to predict chronic kidney disease could potentially solve this issue. We are also developing a feature-based deep learning retinal vessel algorithm for predicting chronic kidney disease. Fourth, we have not performed implementation and cost-effectiveness studies. While dilating pupils improves the quality of images, non-mydriatic images are taken for logistical reasons in many diabetic retinopathy screening programmes. We are developing an image quality algorithm that identifies poor images and alerts the need for repeated retinal photography. Finally, because our study used Asian data (Chinese, Indian, and Malay), our DLA would be particularly useful for Asian countries such as China and India (with a high burden of chronic kidney disease), but further testing in other populations and in more diverse clinical settings will allow the DLA to achieve even more robust performance.

In conclusion, our study shows the potential of a DLA using retinal images to identify chronic kidney disease in community populations. Since access to digital retinal photography is increasing at the community and primary care level, a retinal image-based DLA has the potential to be adopted for first stage chronic kidney disease screening before confirmatory tests.

**Declaration of interests**

TYW reports consultancy and payments for lectures from Allergan, Bayer, Boehringer Ingelheim, Denentina, Merck, Novartis, Oxurion, and Chronic Kidney Disease Screening could be combined with diabetic retinopathy screening in diabetic patients, potentially increasing the chronic kidney disease screening at population level.
Roche, and Samsung, outside the submitted work. All other authors declare no competing interests.

Data sharing
As the study involves human participants, the data cannot be made freely available in the manuscript, the appendix, or a public repository because of ethical restrictions. Nevertheless, the data are available from the Singapore Eye Research Institutional Ethics Committee for researchers who meet the criteria for access to confidential data. Interested researchers can send data access requests to the Singapore Eye Research Institute at seri@seri.com.sg.

Acknowledgments
This study was supported by the National Medical Research Council, OFLCG/001/2017, NMRC/STaR/003/2008, NMRC/0796/2003, and NMRC/STaR/006/2013 and NMRC/CIRC/14/15/2015. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References
1. Fox CS, Matsushita K, Woodward M, et al. Associations of kidney disease measures with mortality and end-stage renal disease in individuals with and without diabetes: a meta-analysis. Lancet 2012; 380: 1662–73.
2. Ene-Iordache B, Perico N, Bikbov B, et al. Chronic kidney disease and cardiovascular risk in six regions of the world (ISN-KDDC): a cross-sectional study. Lancet Glob Health 2016; 4: e307–19.
3. Berns JS. Routine screening for CKD should be done in asymptomatic adults—selectively. Clin J Am Soc Nephrol 2014; 9: 1988–92.
4. Quesenberry CP, Hopkins RH, Sweet DE, et al. Screening, monitoring, and treatment of stage 1 to 3 chronic kidney disease: a clinical practice guideline from the American College of Physicians. Ann Intern Med 2013; 159: 835–47.
5. Manski-Nankervis JE, Thuraissigam S, Lau P, et al. Screening and diagnosis of chronic kidney disease in people with type 2 diabetes attending Australian general practice. Aust J Prim Health 2018; 24: 280–86.
6. Reuters AT. Epidemiology of diabetic kidney disease. Med Clin North Am 2013; 97: 1–18.
7. Wong CW, Wong TY, Cheng CY, Sabanayagam C. Kidney and eye diseases: common risk factors, etiological mechanisms, and pathways. Kidney Int 2014; 85: 1290–102.
8. Lim LS, Cheung CY, Sabanayagam C, et al. Structural changes in the retinal microvasculature and renal function. Invest Ophthalmol Vis Sci 2011; 54: 2970–76.
9. Sabanayagam C, Shankar A, Koh D, et al. Retinal microvascular caliber and chronic kidney disease in an Asian population. Am J Epidemiol 2008; 169: 625–32.
10. Guo Han V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016; 316: 2402–10.
11. Ting DSW, Cheung CY, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA 2017; 318: 2211–23.
12. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017; 542: 115–18.
13. Kuo CC, Chang CM, Liu KT, et al. Automation of the kidney function prediction and classification through ultrasound-based kidney imaging using deep learning. NPJ Digit Med 2019; 2: 20.
14. Foong AW, Saw SM, Loo JL, et al. Rationale and methodology for a population-based study of eye diseases in Malay people: the Singapore Malay eye study (SiMES). Ophthalmic Epidemiol 2007; 14: 25–35.
15. Lavanya R, Jegannathan VS, Zheng Y, et al. Methodology of the Singapore Indian Chinese Cohort (SICC) eye study: quantifying ethnic variations in the epidemiology of eye diseases in Asians. Ophthalmic Epidemiol 2009; 16: 325–36.
16. Sabanayagam C, Yip W, Gupta P, et al. Singapore Indian Eye Study-2: methodology and impact of migration on systemic and eye outcomes. Clin Exp Ophthalmol 2017; 45: 779–89.
17. Sabanayagam C, Tai ES, Shankar A, Lee J, Sun C, Wong TY. Retinal arteriolar narrowing increases the likelihood of chronic kidney disease in hypertension. J Hypertens 2009; 27: 2209–17.
18. Xu J, Xu L, Wang YX, You QS, Jonas JB, Wei WB. Ten-year cumulative incidence of diabetic retinopathy. The Beijing Eye Study 2001/2011. PLoS One 2014; 9: e111320.
19. Levey AS, Stevens LA, Schmid CH, et al. A new equation to estimate glomerular filtration rate. Ann Intern Med 2009; 150: 604–12.
20. Xu D, Lee ML, Wynne Hsu. Propagation mechanism for deep and wide neural networks. 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Long Beach, CA, USA; June 16–20, 2019; 9212–20.
21. Saunders MR, Cifu A, Vela M. Screening for chronic kidney disease. JAMA 2015; 314: 615–16.
22. Mitka M. Nephrologists question ACP’s kidney disease guidelines. JAMA 2013; 310: 2387–88.
23. National Kidney Foundation. K/DOQI clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratification. Am J Kidney Dis 2002; 39 (2 suppl 1): S1–266.
24. Hallan SI, Dahl K, Oien CM, et al. Screening strategies for chronic kidney disease in the general population: follow-up of cross sectional health survey. BMJ 2006; 333: 1047.
25. Komenda P, Ferguson TW, Macdonald K, et al. Cost-effectiveness of primary screening for CKD: a systematic review. Am J Kidney Dis 2014; 63: 789–97.
26. Bang H, Vupputuri S, Shoham DA, et al. Screening for Occult RENal Disease (SCORED): a simple prediction model for chronic kidney disease. Arch Intern Med 2007; 167: 374–81.
27. Bradshaw C, Kondal D, Montez-Rath ME, et al. Early detection of chronic kidney disease in low-income and middle-income countries: development and validation of a point-of-care screening strategy for India. BMJ Glob Health 2019; 4: e001646.
28. Delgado R, Tibau XA. Why Cohen’s Kappa should be avoided as a performance measure in classification. PLoS One 2019; 14: e0222916.
29. Ozenne B, Subtil F, Maucourt-Boulch D. The precision-recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. J Clin Epidemiol 2015; 68: 855–59.
30. Deacon B, Abramowicz J. Fear of needles and vasovagal reactions among phlebotomy patients. J Anxiety Disord 2006; 20: 946–60.