A Systematic Review and Meta-Analysis of Applying Deep Learning in the Prediction of the Risk of Cardiovascular Diseases From Retinal Images

tvst

Review

A Systematic Review and Meta-Analysis of Applying Deep Learning in the Prediction of the Risk of Cardiovascular Diseases From Retinal Images

Wenyi Hu[1,2], Fabian S. L. Yii[3,4], Ruiye Chen[1,2], Xinyu Zhang[5], Xianwen Shang[1], Katerina Kiburg[1], Ekaterina Woods[1], Algis Vingrys[1,6], Lei Zhang[7], Zhuoting Zhu[1], and Mingguang He[1,2]

[1] Centre for Eye Research Australia, Royal Victorian Eye and Ear Hospital, East Melbourne, Australia
[2] Department of Surgery (Ophthalmology), The University of Melbourne, Melbourne, Australia
[3] Centre for Clinical Brain Sciences, Edinburgh Medical School, University of Edinburgh, Edinburgh, UK
[4] Curle Ophthalmology Laboratory, Institute for Regeneration and Repair, University of Edinburgh, Edinburgh, UK
[5] Shanghai Jiaotong University School of Medicine, Shanghai, China
[6] Department of Optometry and Vision Sciences, The University of Melbourne, Melbourne, Australia
[7] Central Clinical School, Faculty of Medicine, Nursing and Health Sciences, Monash University, Melbourne, Australia

Correspondence: Mingguang He, Department of Surgery (Ophthalmology), The University of Melbourne, Melbourne, Australia, 32 Gisborne Street, East Melbourne, Victoria 3002, Australia. e-mail: mingguang.he@unimelb.edu.au
Zhuoting Zhu, Centre for Eye Research Australia, Royal Victorian Eye and Ear Hospital, 32 Gisborne Street, East Melbourne, Victoria 3002, Australia. e-mail: lisa.zhu@unimelb.edu.au
Lei Zhang, Central Clinical School, Faculty of Medicine, Nursing and Health Sciences, Monash University, 37, Rainforest Walk, Clayton, Victoria 3168, Australia. e-mail: lei.zhang1@monash.edu
Algis Vingrys, Department of Optometry and Vision Sciences, The University of Melbourne, 800, Swanston Street, Melbourne, Victoria 3053, Australia. e-mail: algis@unimelb.edu.au

Received: March 24, 2023
Accepted: June 8, 2023
Published: July 13, 2023

Keywords: artificial intelligence (AI); cardiovascular diseases (CVDs); retinal Imaging

Purpose: The purpose of this study was to perform a systematic review and meta-analysis to synthesize evidence from studies using deep learning (DL) to predict cardiovascular disease (CVD) risk from retinal images.

Methods: A systematic literature search was performed in MEDLINE, Scopus, and Web of Science up to June 2022. We extracted data pertaining to predicted outcomes, model development, and validation and model performance metrics. Included studies were graded using the Quality Assessment of Diagnostic Accuracies Studies 2 tool. Model performance was pooled across eligible studies using a random-effects meta-analysis model.

Results: A total of 26 studies were included in the analysis. There were 42 CVD risk-related outcomes predicted from retinal images were identified, including 33 CVD risk factors, 4 cardiac imaging biomarkers, 2 CVD risk scores, the presence of CVD, and incident CVD. Three studies that aimed to predict the development of future CVD events reported an area under the receiver operating curve (AUROC) between 0.68 and 0.81. Models that used retinal images as input data had a pooled mean absolute error of 3.19 years (95% confidence interval [CI] = 2.95–3.43) for age prediction; a pooled AUROC of 0.96 (95% CI = 0.95–0.97) for gender classification; a pooled AUROC of 0.80 (95% CI = 0.73–0.86) for diabetes detection; and a pooled AUROC of 0.86 (95% CI = 0.81–0.92) for the detection of chronic kidney disease. We observed a high level of heterogeneity and variation in study designs.

Conclusions: Although DL models appear to have reasonably good performance when it comes to predicting CVD risk, further work is necessary to evaluate the real-world applicability and predictive accuracy.

Translational Relevance: DL-based CVD risk assessment from retinal images holds great promise to be translated to clinical practice as a novel approach for CVD risk assessment, given its simple, quick, and noninvasive nature.

translational vision science & technology

Citation: Hu W, Yii FSL, Chen R, Zhang X, Shang X, Kiburg K, Woods E, Vingrys A, Zhang L, Zhu Z, He M. A systematic review and meta-analysis of applying deep learning in the prediction of the risk of cardiovascular diseases from retinal images. Transl Vis Sci Technol. 2023;12(7):14, https://doi.org/10.1167/tvst.12.7.14

Introduction

Although cardiovascular diseases (CVDs) are the leading cause of mortality globally, up to 80% of premature CVD events can be prevented by managing modifiable risk factors through lifestyle improvement.[1,2] A comprehensive risk assessment considering multiple risk factors is effective in preventing CVD events[3] and is therefore considered the first key step for early identification and intervention of high-risk patients.[4] Numerous risk assessment tools that integrate various clinical risk factors have been developed in the past, such as the Framingham risk score, Systemic COronary Risk Evaluation (SCORE), and QRISK.[5–7] Although these risk-factor-based assessment tools are well-established and even adopted by guidelines, the involvement of blood test results and the limited time for comprehensive data collection and entry in real-world clinics may limit their widespread usage.[8]

The retina is recognized as a "window" to visualize and assess cardiovascular health noninvasively,[9] as it is considered to share similar anatomic structure and physiological function with cardiac vasculature.[9] Previous studies have indicated associations between various retinal features and the risk of developing CVD, ranging from retinal vascular geometry/morphology (i.e. vessel caliber, branching angle, tortuosity, and fractal dimension), retinal vascular network patterns, and retinal pathologies (cotton wool spots, arteriovenous nicking, and microaneurysm).[10–13] However, exclusively focusing on specific measurements might overlook some implicit information and underestimate the potential of the retina as a whole to inform cardiovascular (CV) health.

Deep learning (DL) is a subfield of the artificial intelligence (AI) techniques that focuses on utilizing artificial neural networks with multiple computational layers to learn and extract complex predictive features from high-dimensional data, including medical images.[14] Recent studies have developed DL algorithms that could make accurate diagnoses for diseases that are largely dependent on morphology, such as diabetic retinopathy, returning comparable or slightly better performance from AI than human graders.[15]

With all these resources, it is an emerging area of research to investigate the application of DL to predict the risk of CVD using retinal images. However, distinct datasets and methodologies were adopted in the development and validation of the algorithms, and a variety of CVD-related risk factors or outcomes were used as the prediction outcome in different studies. Recently, Wong et al. have conducted a narrative review on the current scope and future directions of AI on retinal images for CVD prediction.[16] Arnould et al. also conducted a review focusing on the application of retinal vascular networks and the development of oculomics in future CVD risk assessment.[17] Despite the importance of these studies, very few studies have performed a systematic review to comprehensively evaluate the characteristics of these studies.

Therefore, we aim to perform a systematic review and meta-analysis to evaluate the studies that apply DL and retinal images in the prediction of CVD risk, to comprehensively understand the characteristics of these studies in terms of model development and validation, predictors, and CVD-risk-related outcomes and quantify the performances of the DL models in the prediction of CVD risk.

Methods

We conducted the systematic review following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist.[18] This study protocol is registered on the International Prospective Register of Systematic Reviews (PROSPERO) under the ID CRD42022364921.

Literature Search

We performed a systematic literature search in PubMed, Scopus, and Web of Science on June 30,

2022, using search terms to identify eligible articles with no limitations on the publication year. The keywords used for the search included "artificial intelligence," "machine learning," or "deep learning," combined with "retinal image," "retinal photo," "fundus photo," "fundus image," and combined "cardiovascular disease," "cardiovascular risk," "coronary heart disease," "myocardial infarction," or "stroke," and "risk assessment" or "screening." The detailed search strategy is shown in the Appendix 1. Titles and abstracts were independently reviewed by two researchers (authors W.H. and R.C.), and all relevant citations were included for full-text analysis.

Eligibility Criteria

Studies that reported the performance of DL algorithms in predicting CVD risk, including CVD risk factors, CVD risk score, or incident CVD events, and using retinal images as their input were included. Studies that reported DL algorithms using vascular segmentation or extracting specific retinal features from the retinal image to predict CVD risk were excluded from this study as the present review focused on the DL pipeline that took a whole image directly as input, which provides a more holistic approach to risk prediction by understanding different predictive information from all potential retinal features and arguably has greater clinical utility insofar as automated prediction is concerned. Included studies had to be written in English, conducted on human subjects, and report original research. Preprints and conference papers were also included. To find any other potentially relevant research, relevant studies cited by eligible studies were also examined.

Data Extraction and Quality Assessment

Data extraction and quality assessment were performed by one reviewer (authors W.H.) and underwent double check by another two reviewers (authors X.Z. and F.Y.) independently. Relevant data were extracted from the retrieved articles using a predefined Excel spreadsheet. For all the retrieved articles, the extracted data included the first author, publication year, type of DL model structure, predictive horizon, the study populations, the input(s) for the development of the DL models (retinal images with/without other inputs), the reference standard (the prediction outcome of the DL models), the datasets used, sample sizes, and the number of outcomes for model development and internal validation, the type of internal validation (cross-validation or random split), and external validation if available. We also included the measurement and results used in assessing the diagnostic accuracy or reliability of the DL algorithms in the internal and external validation datasets, and the features highlighted in the retrieved attention maps of the algorithms if available.

In addition, more data were extracted for a subset of studies of those that the prediction outcomes were incident CVD cases. The extracted data included basic article information, the definition of CVD events (including the prediction horizon of incident CVD), follow-up periods, predictors (retinal images per se or retina-predicted intermediate traits), sample sizes, and cases of the cohorts, modeling method, and the measurement adopted with results of diagnostic accuracy.

Quality assessment was performed during data extraction, using Quality Assessment of Diagnostic Accuracies Studies 2 (QUADAS-2).[19] This assessment tool designed for diagnostic accuracy studies systematically evaluated the risk of bias in four domains, including patient selection, index test, reference test, and flow and timing, and concerns of the applicability of the studies regarding the review questions in all the domains except for flow and timing.

Data Synthesis and Analysis

Data synthesis was performed qualitatively to summarize the characteristics of included studies, including the type of DL model structures used, the CVD risk-related prediction outcomes, the predictors and sample size of studies, the studies that underwent external validation, and the studies that generated attention maps and the features highlighted in different studies for certain CVD risk-related prediction outcomes.

Studies with the same predictor (retinal images or retinal images with clinical data), prediction outcome, and measure(s) of diagnostic accuracy or reliability, performed in two or more studies with three or more cohorts were synthesized for meta-analysis. Random-effect models were used to estimate the pooled effect sizes across studies, using precomputed effect sizes and their 95% confidence intervals (CIs) extracted from the articles. If the algorithm was evaluated in both internal validation and external validation in different cohorts, it was considered as a different study in the meta-analysis. I^2 was used to evaluate the heterogeneity between studies (25%–50%, 50%–75%, and 75% and over represent low, moderate, and high heterogeneity, respectively). All statistical analyses were performed using Stata version 16 (StataCorp LLC, College Station, TX).

Results

A total of 859 records were found using the search strategy, and 277 duplicates were removed, resulting in a total of 582 records for screening. Five hundred thirty-eight records were excluded after reviewing the titles and abstracts. Of the remaining articles for full-text screening, 18 studies were excluded because of the adoption of vessel segmentation or feature extraction instead of whole retinal image analysis as the input, using other irrelevant imaging modalities or irrelevant prediction outcomes, being a case study, without sufficient data on the datasets and outcome measures, or being retracted. A total of 26 studies were included in the study for qualitative appraisal.[20–45] Three studies were synthesized for the pooled performance of the algorithms used for age prediction, four for gender prediction, three for diabetes detection, and two for chronic kidney disease (CKD) detection. Results from internal and external validation cohorts, if available, of the same study were included for the meta-analysis. The PRISMA flow diagram is shown in Figure 1. The data extracted to demonstrate the characteristics of all the included studies are shown in detail in Appendix 2. The characteristics of a subset of these studies that utilized retinal images or retinal image-predicted intermediate traits to further predict incident CVD events are listed separately in Appendix 3. The quality assessment results are displayed in Figure 2 and details are described in Appendix 4. The main reasons for the risk of bias lie in the adoption of a case-control designs in

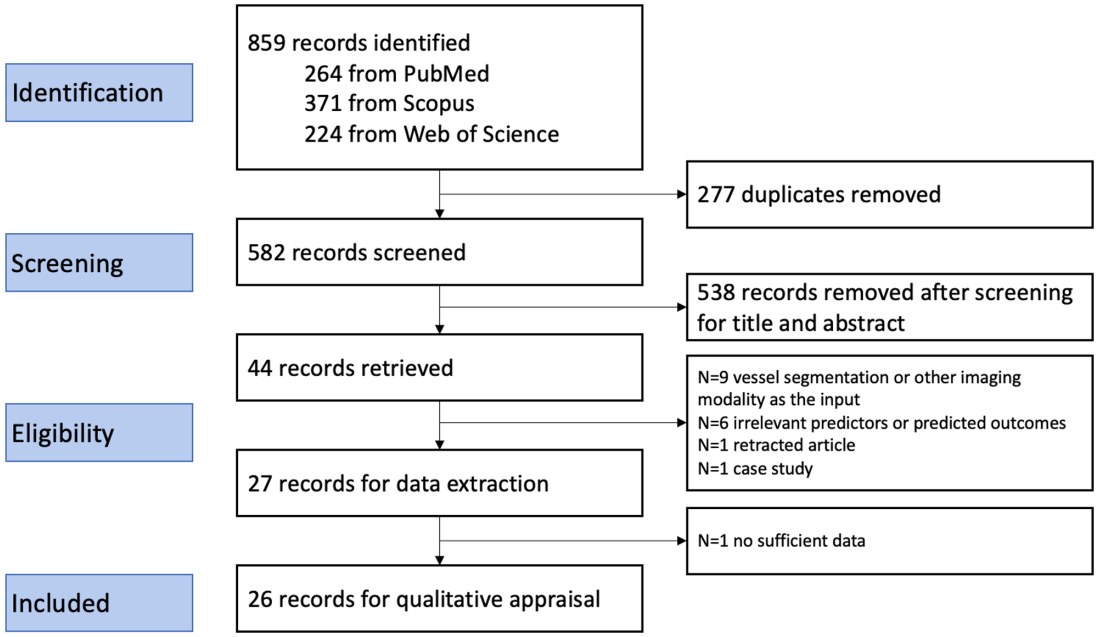

Figure 1. PRISMA flow diagram.

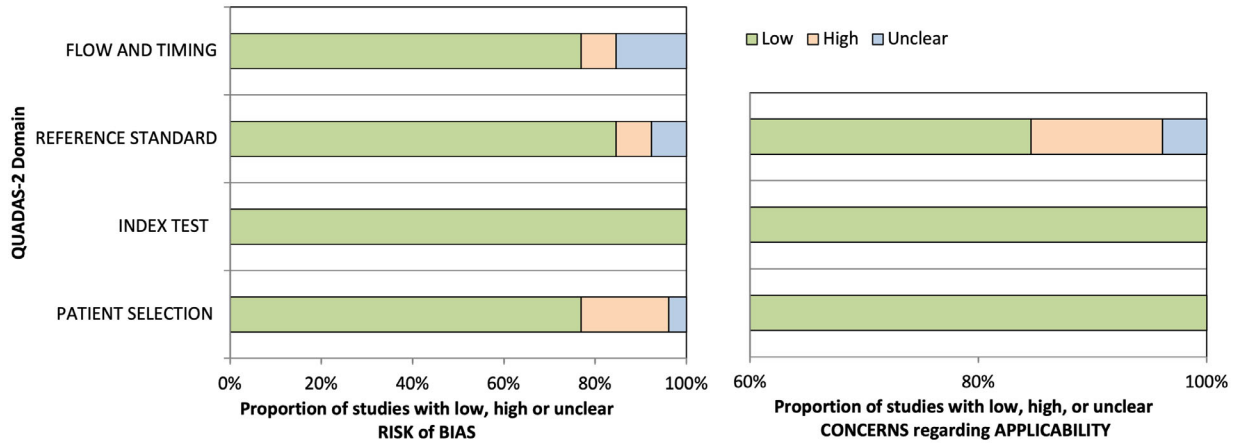

Figure 2. Diagram of quality assessment results of the 26 included studies using QUADAS-2.

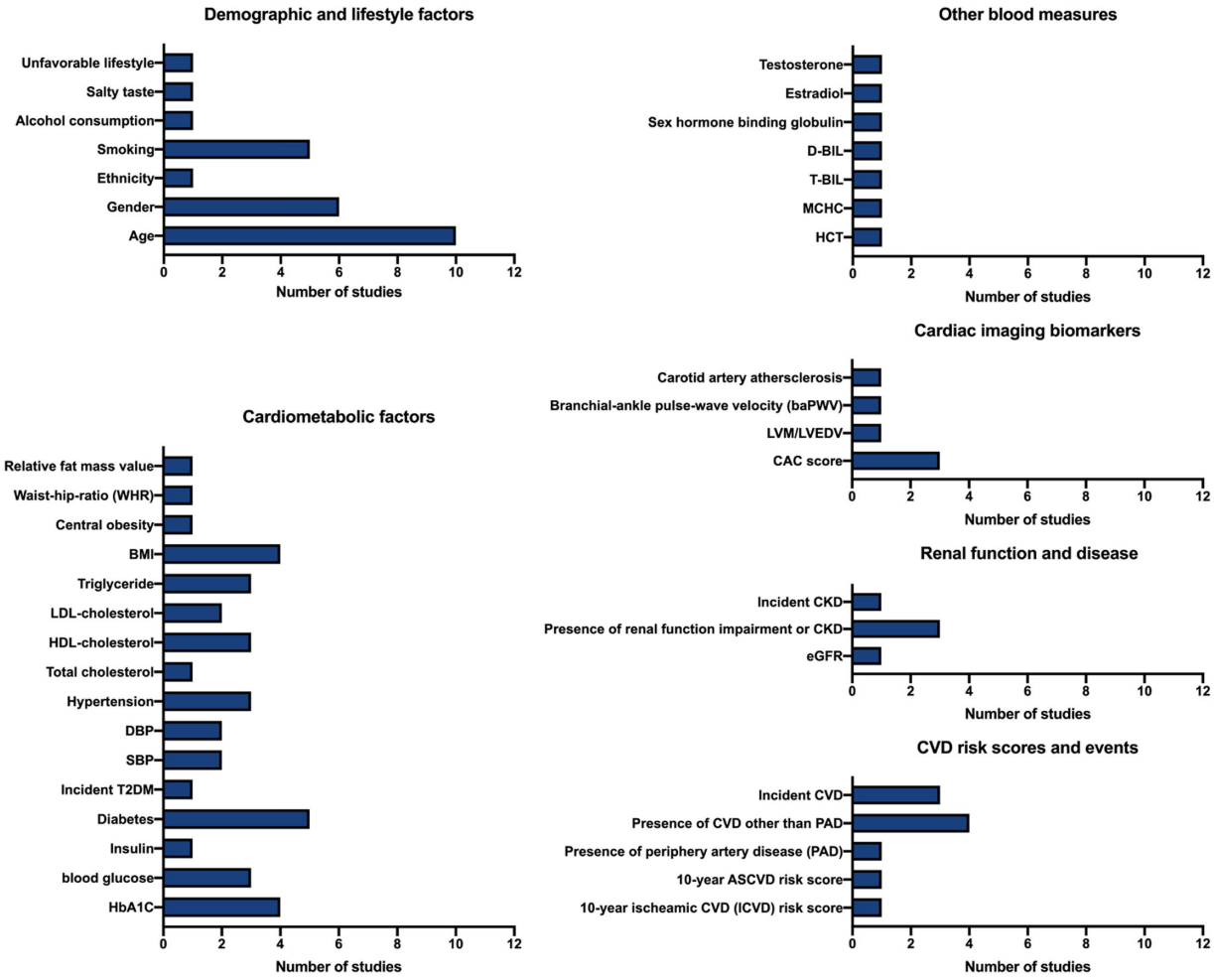

Figure 3. Summary of CVD risk-related prediction outcomes in the included studies.

terms of patient selection, the unclear definition of the reference standards, and the use of inappropriate intervals between the index test (retinal image acquisition) and the reference standard.

CVD Risk-Related Outcomes

A total of 42 CVD risk-related outcomes were predicted using retinal images in the included studies as some studies predicted more than one outcome. The detailed breakdown of the CVD risk-related outcomes that have been used for prediction from retinal images are depicted in Figure 3. A total of 33 of them were CVD risk factors spanning the following categories: demographic factors and lifestyle, cardiometabolic factors, other blood measures, renal function, and CKD. Four of them are cardiac imaging biomarkers. The other CVD risk-related outcomes included CVD risk scores (10-year ischemic CVD [ICVD] risk score) and 10-year atherosclerotic CVD [ASCVD] risk score), presence of CVD (presence of CVD and presence of peripheral artery disease, and presence of stroke), and incident CVD. Among these, 6 studies predicted the longitudinal CVD risk-related outcomes, including incident type 2 diabetes, incident CKD, and incident CVD. The most common CVD risk-related outcomes were: age ($n = 10$), gender ($n = 6$), smoking status ($n = 5$), diabetes ($n = 5$), HbA1C ($n = 4$), body mass index (BMI; $n = 4$), blood glucose ($n = 3$), hypertension ($n = 3$), HDL-cholesterol ($n = 3$), coronary artery calcium (CAC) score ($n = 3$), presence of renal impairment or CKD ($n = 3$), presence of CVD ($n = 4$), and incident CVD ($n = 3$). Notably, different definitions of the reference standards may be used in different studies, which can be found in detail in Appendix 2.

Model Inputs, Development, and Internal Validation

Retinal images as the only input for the development of DL models were investigated in 25 (96.2%)

studies, and one study only evaluated the multi-modal DL algorithm with both retinal images and clinical metadata as the input. Three studies experimented with the used of retinal images only and multi-modal input to assess the performance of DL models in the prediction of CVD risk-related outcomes. In the three studies that predicted incident CVD events, Poplin et al. used retinal images per se (and clinical metadata) for prediction,[26] whereas Rim et al. and Diaz.Pinto et al. used retina-predicted endophenotypes, including the RetiCAC (i.e. retina-predicted probability of the presence of CAC) with demographic factors and retina-predicted left ventricular mass (LVM)/left ventricular end-diastolic volume (LVEDV) with demographic factors, respectively.[22,24]

A variety of DL algorithms were used for model development in the studies, including Convolutional Neural Network (CNN) architectures, such as visual geometry graphic (VGG), Inception, ResNet, Inception-ResNet, Xception, EfficientNet, DiaNet, MobileNet, cCondenseNet, DenseNet, NASNet-Large, and one CNN architecture not specified. The CNN architectures were also used in combination with other models. For example, Mueller et al. utilized multiple instance learning (MIL) to process images to reserve high resolution of the input retinal images.[29] Diaz Pinto et al. applied a multichannel variational autoencoder (mcVAE), a multimodal AI model, that integrates cardiac magnetic resonance imaging (MRI) scans and retinal images for CVD risk prediction.[24]

Different terms were used in distinct studies for the development, internal validation, and external validation of the DL model. In this review, development datasets included both training and tuning sets in the original articles. Internal validation in this review was called the internal test in some studies. The sample sizes of the development datasets were reported in terms of retinal images and/or individuals, which ranged between 135 images from 77 participants and 798,866 images from 390,947 participants. Seven studies that predicted a categorical outcome did not specify the number of cases included in the development datasets. Two of them reported the cases in the internal validation set and the number can be estimated given the random split of samples, whereas the other five studies specified the number of cases neither in the development nor internal validation set.

There were 18 (69.2%) studies that randomly split the total sample for development and validation, whereas 8 (30.8%) studies used k-fold cross-validation for the development and internal validation of the DL algorithms. The details of the DL model performances of the internal validation can be found in detail in Appendix 2.2. Regarding age prediction from retinal images, the DL algorithms in different studies achieved a mean absolute error (MAE) ranging from 2.74 to 3.55 years.[26,31,32,34,42] For gender prediction, the AUROC of the DL algorithms was between 0.704 and 0.978 among all relevant studies.[26,28,31,32,40,44] Among the studies that investigated the prediction of smoking status, the models achieved an AUROC that ranged from 0.71 to 0.86.[26,28,30,32,44] The DL models had an MAE between 0.61% and 1.39% for the prediction of HbA1c value[26,32] and an MAE between 0.652 and 1.06 mmol/l for the prediction of blood glucose level.[32,38] Among the studies that predicted the presence of diabetes, the models had an AUROC ranging from 0.731 to 0.923 using retinal images only, which increased to 0.929 by adding clinical data.[23,28,33,38,40] The MAE ranged from 8.96 to 11.35 mm Hg for systolic blood pressure prediction, 6.42 to 6.84 mm Hg for diastolic blood pressure, and 3.29 to 4.31 kg/m^2 for BMI prediction.[26,32] The AUROC for prediction of prevalent CKD was between 0.911 and 0.918 and was up to 0.938 when clinical metadata was added to the model.[37,38] As for the prediction of prevalent CVD, the AUROC ranged from 0.499 to 0.700,[40,44,45] and the accuracy was reported to be 0.756 when retinal images were used only and increased to 0.783 when dual-energy X-ray absorptiometry (DXA) data was added.[39]

External Validation

Six (23.1%) studies validated the performance of the DL algorithms using a total of nine external validation sets. Sample sizes of the external validation sets ranged from 1054 images from 527 participants to 56,301 participants. The six studies were developed in Korean ($n = 2$), Chinese ($n = 2$), multi-ethnic (Malay, Chinese, and Indian, $n = 1$), and Caucasian ($n = 1$) populations. Only two of the studies were validated in a population of different ethnicities. Nusinovici et al.[25] developed the DL algorithm for the prediction of age in the Korean population and validated the performance in the UK Biobank, which predominantly consisted of Caucasians. The AUROC of predicting the probability of being 65 years and over was 0.968 (95% CI = 0.965–0.970) in the internal validation set, and 0.756 (95% CI = 0.753–0.759) in the external validation set. Yun et al.[40] developed an algorithm for the detection of type 2 diabetes in the UK Biobank population, which was validated in a sample composed of Korean patients with diabetes and non-diabetic patients from the UK Biobank. The AUROC in the internal and external validation set was 0.731 (95% CI = 0.707–0.756) and 0.703 (95% CI = 0.691–0.715), respectively.

Performance of the AI Algorithm and Meta-Analysis

The performance of the DL algorithms was measured in a series of parameters, including the diagnostic accuracy and reliability measures such as sensitivity, specificity, AUROC, accuracy positive predictive value (PPV), negative predictive value (NPV), area under the precision-recall curve (AUPRC), F1 score for binary prediction outcomes, MAE, and the limits of agreement for continuous outcomes. The details of the performance measures were listed by prediction outcome in Appendix 2.2. Meta-analysis was performed for eligible studies that predicted age, gender, diabetes, and the presence of CKD. The detailed results are shown in Appendix 5.

CVD Risk Score Prediction

Two studies investigate the prediction of CVD risk scores from retinal images only. Ma et al.[35] developed and validated the algorithm to predict a 10-year ICVD risk score, which was defined by 7 parameters determined by the Cox regression models, including age, sex, systolic blood pressure, total cholesterol, BMI, smoking status, and diabetes. The AUROC for borderline risk (ICVD risk >5%) was 0.971 (95% CI = 0.967–0.975) and the AUROC for intermediate or high risk (ICVD risk >7.5%) was 0.976 (95% CI = 0.973–0.980) on an individual level. The R^2 was 0.876, indicating that 87.6% of the variability in the retina-predicted score is explained by the ICVD risk score. Syed et al.[41] performed a study to predict a 10-year ASCVD risk score, which was defined by the pooled cohort equations (PCEs) consisting of age, sex, systolic and diastolic blood pressure, total and high-density lipoprotein cholesterol, diabetes, and smoking status. The MAE was 0.1085 (96% CI = 0.1053, 0.1116) and R^2 was 0.5338 (95% CI = 0.5036, 0.5628).

Longitudinal Prediction of Incident CVD Events

Three studies investigated the prediction of incident CVD events longitudinally (Appendix 3). Poplin et al.[26] developed an algorithm using retinal images per se to predict 5-year major adverse CVD events in the UK Biobank study, achieving an AUROC of 0.70 (95% CI = 0.65–0.74). When adding well-established clinical risk factors or SCORE to the model, the AUROCs were 0.73 (95% CI = 0.69-0.77) and 0.72 (95% CI = 0.67–0.76), respectively. Rim et al.[22] tested the predictive value of RetiCAC, which was an intermediate trait predicted from retinal images, representing the probability of CAC presence. The combination of RetiCAC with age and gender, or with PCE was modeled to predict incident fatal (and non-fatal) CVD events during a median follow-up period ranging from 4.1 years to 10.3 years in multiple cohorts. The C-statistics ranged between 0.68 (95% CI = 0.58–0.79) (retinal images and age and gender in the CMERC-HI study) and 0.806 (95% CI = 0.790–0.828; retinal images and PCE in the SEED study). Diaz Pinto et al.[24] used another endophenotype predicted from the retinal image, namely the retina-predicted LVM and retinal-predicted LVEDV in combination with minimal demographic factors to predict incident myocardial infarction. The mean (± standard deviation) accuracy was 0.74 ± 0.03 and the mean (± standard deviation) AUROC was 0.8 ± 0.02 in the UK Biobank from which the algorithm was developed. The AUROC and accuracy were 0.59 and 0.59 when tested in the age-related eye disease study (AREDS) with all age-related macular degeneration (AMD) images included and 0.70 and 0.68 when all AMD images were excluded.

Attention Maps

Attention maps were retrieved in 17 (65.4%) studies. For the prediction of age, the features highlighted included the macula, optic disc (including optic nerve head), blood vessels (including vessel arcade at the posterior pole), and regions around the blood vessels.[25–27,31,34] Features in gender prediction comprised of optic discs, macula (including the fovea), and vessels.[26,31] Specifically, Kim et al. reported that the proximal vascular regions were prominently highlighted in women.[31] Blood vessels, perivascular regions, and the fovea were indicated in the attention maps in the prediction of smoking status.[26,30] Nonspecific perivascular surroundings were featured in one study predicting HbA1C.[26] In the prediction of diabetes, the central retina area between the optic disc and the macula was highlighted in one study.[23] In addition, another study found features scattered throughout the whole image that may correspond to diabetic retinopathy (DR) changes in some cases, in patients with diabetes with/without DR.[38] Blood vessels and nonspecific features were highlighted in the prediction of systolic blood pressure and diastolic blood pressure, respectively.[26] Nonspecific features were found for the prediction of BMI as well.[26] Blood vessels, in specific main retinal branches, vessels around the optic disc and vascular arcades, the optic nerve head and its pathologic changes, and retinal pathologies, such as cotton-wool spots were the major features highlighted for the prediction of cardiac imaging biomarkers, including CAC

score, branchial-ankle pulse wave velocity (baPWV), and carotid atherosclerosis.[21,22,27,43] As for the prediction of renal function impairment or CKD, retinopathy changes, retinal vascular geometry/morphology (including venous caliber, vessel density, vessel branch points, and arterio-venous junctions) and the optic nerve were featured.[36–38] Optic disc, vessels, and macula were extracted in the prediction of CVD risk score.[41] The central region of the retina, microhemorrhage, vessel arcades, and the optic disc were documented in the studies that predicted the presence of CVD.[29,39]

Discussion

This study was conducted to qualitatively appraise the characteristics of the studies that apply DL to predict CVD risk-related outcomes from retinal images and to quantify the diagnostic accuracy of studies that were designed to predict the same CVD risk-related outcomes. This systematic review identified a variety of CVD risk-related outcomes that were predicted from retinal images and the diagnostic accuracy of commonly predicted CVD risk-related outcomes, such as age, gender, diabetes, and CKD, were highly acceptable in the experimental settings. Nevertheless, the results need to be interpreted with caution given the limited number of studies and the between-study heterogeneities.

The application of DL to predict CVD risk from the retinal image is an emerging area of research at the exploratory stage. In terms of prediction outcome of interest, more than 40 CVD risk-related outcomes were summarized in this study. Regarding the performances of the DL algorithms, variations can be found in the performances of the DL algorithms that predicts certain CVD risk-related outcome among different studies. The differences may lie in the variation of sample sizes, different DL algorithms used, and different populations involved. This suggests that even the algorithms achieved a reasonably good performance in its development and validation, it warrants further validation in different datasets to increase the reliability, robustness, and generalizability. In the limited number of studies that are eligible to examine the pooled performance of the algorithms, the results showed that the DL models achieved an area under the curve (AUC) of at least 0.80 in discriminating gender, diabetes status, and CKD. However, although algorithms reported in the study achieved an acceptable performance in predicting single risk factors, it should be borne in mind that some of the risk factors are nonspecific if used alone, and some are intuitive without the need for retinal images and analytical algorithms. In terms of the prediction scheme of the CVD risk-related outcomes, most studies are detecting the CVD risk-related features cross-sectionally, and the prediction of future risk is not defined or further investigated. There could be difficulties in the accurate prediction of future CVD events as the effects of treatment and future comorbid changes will become manifest over a longitudinal period.[46] This may explain why the majority of prediction schemes for cross-sectional endophenotypes or CVD risk scores are retinal images only, whereas the prediction of future CVD risk requires the addition of clinical risk factors to improve the performance of the AI algorithms.

In terms of study reporting, some studies put more emphasis on algorithm development and did not describe in detail the selected datasets, the definition of the reference tests, the image acquisition protocols, and the flow and timing of data collection, which may potentially introduce biases and limit the validity of the studies from a clinical perspective of view. The performance of the DL algorithms was also reported in a variety of measurements. For some categorical outcomes, only AUC or accuracy was reported but no sensitivity, specificity, PPV, and NPV were displayed, which might limit a holistic interpretation of the algorithms. Further, as different measurements will provide barriers to a pooled analysis and further translation of the technology in terms of establishing clinical cutoff values or criteria.[47] This could be addressed in the future by developing and promoting the utilization of standard reporting checklists for these trials, such as Standards for Reporting Diagnostic accuracy studies - Artificial Intelligence (STARD-AI).[48]

The generalizability of the studies should be taken into consideration. Most studies were developed and validated in the Caucasian and Asian populations, and there is scarce evidence for the other ethnicities that may be at a higher risk of CVD, such as African Americans, aboriginal people in Canada, or indigenous Australians.[49–52] The performance of the DL algorithms reported in the studies might consider measures that do not exist or are inappropriate for these populations thereby over- or underestimating their importance so the robustness of such indices needs to be interpreted with caution. Moreover, a primary challenge for ensuring the generalizability of DL models arises from domain shift, referring to the differences in the distributions of the data between the training sets and test sets, such as clinical practical settings. This shift may be due to the distinct imaging protocols or machines provided by vendors or different characteristics of the patient populations.[53]

In addition, the DL algorithms will perform worse on a new dataset compared with the original one, because of overfitting or data leakage.[53] Only a small proportion of studies in the systematic review were externally validated. In addition, for all the studies included in the analysis, retrospective analyses of existing datasets were applied for development and validation. This could be attributable to the challenge that large numbers of images need to recruited for DL algorithm development in cases where prospective data collection are used. This is time-consuming and costly. Finally, as retrospective analysis allows image quality assessment in advance and exempts the consideration of the image acquisition success rate, it can be expected that the application of these prototypes prospectively in real-world clinical settings will face greater challenges.

Nevertheless, the application of DL and retinal images in predicting CVD risk shows significant clinical implications. Compared to risk factor-based models, the use of DL and retinal images may create a more efficient and labor-free approach for CVD risk assessment. The traditional parametric prediction models combining multiple risk factors that were considered to be cumbersome in real-world practice.[8] The retina-based CVD risk assessment could potentiate efficiency by removing the need for blood tests, and providing extra information on end-organ damage. Retinal photography is one of the most widely adopted and cost-effective imaging modalities used for routine eye care.[54] The development of retinal image acquisition technologies, handheld fundus cameras, or smartphone-based fundus photography will further increase the accessibility of the service.[55,56]

The adoption of DL technology to analyze retinal images has already been tested in a more advanced stage in terms of detecting eye diseases, such as DR, from which the experience can be learnt for the future implementation in terms of CVD risk assessment. Google has conducted a prospective interventional cohort study in Thailand to investigate the performance of AI as a real-time DR screening service in community care settings. For vision-threatening DR, the DL system achieved a comparable accuracy (94.7% vs. 95.3%) and specificity (95.4% vs. 95.5%) to the retinal specialists on-site and higher sensitivity (91.4% vs. 84.8%).[57] Based on evidence from prospective studies that prove the diagnostic accuracy of AI application, recent years have also witnessed the approval of two AI screening systems, EyeArt and IDx-DR by the US Food and Drug Administration (FDA) to detect more-than-mild DR (mtmDR).[58,59] Taken together, the evidence from DR has shed light on the feasibility and the impact of applying automated AI systems can have on retinal image analysis and disease diagnosis. Although no prospective studies have been performed to validate the real-world performance of the AI systems that predict CVD, we can extrapolate the great potential of implementing it in clinical practice from the DR screening experience.

Future studies and collaborations are needed to improve the data source with sufficient sample sizes and upgrade the algorithms to predict CVD events accurately. The constraint of limited sample size could be a major barrier in developing reliable algorithms.[60] A couple of studies examined in this review has fewer than 200 labeled images,[20,27] potentially due to the high intrinsic cost of obtaining certain sufficient data for various labels, which to some extent prevents the DL models to achieve their full potential. Furthermore, the limited number of longitudinal cohort studies with sufficient incident CVD data might be a reason of the scarcity of evidence on retinal images in the prediction of incident CVD. To date, only one study used retinal images as the sole input for future CVD events, with moderate-to-good discrimination of 0.70. The performance is comparable to well-established clinical risk factor-based models, such as the FRS, PCE, SCORE, and QRISK (AUC = 0.63 to 0.82).[61–64] Therefore, future collaborations are needed for the availability of multi-ethnic large-scale longitudinal cohorts with follow-up information on both predictors and outcomes to facilitate the development and validation of algorithms. In addition to the availability of datasets, introducing the concept of open-source algorithms to this field is also a beneficial way to accelerate the translation and customization of the technology as the availability of the codes enables more validation and enhances the performance, facilitates the adjustment of the technology to specific clinical settings, reduces the cost, and enhances benign competition between similar products.[65]

The clinical implication and integration of this technology into a feasible model of care requires careful consideration. Compared to the diagnosis of eye diseases which directly detected the pathologies from the ocular images, CVD risk prediction is more challenging as it is a multifactorial systemic disease, the feature on the retinal images to inform cardiovascular health can be influenced by the complex interplay of environmental factors and genetic risk factors.[66] Therefore, accurate prediction of well-established CVD risk scores might be a solution that could provide extra individualized information than traditional risk factor-based models and simplify the clinical procedure, such as negating invasive tests and complex manual data collection. Multimodal AI models combining retinal images and minimal clinical metadata might be another

solution to improve the prediction performance of the algorithms. But the inclusion of clinical data should take into consideration the principle of noninvasive and simple data acquisition procedures to maximize the benefits of using retinal images as the predictor in real practice over traditional risk factor-based models.

Strengths and Limitations of the Study

This systematic review comprehensively examines the studies using DL to predict CVD risk-related outcomes from retinal images, using structured search terms, literature search, and extracting a wide range of data which enables interpretation of the studies from multiple angles. However, a couple of limitations need to be made aware of. First, there is high heterogeneity in the included studies in terms of CVD risk-related outcomes, methodologies, and measurements of outcomes, which prevented us from performing quantification for all prediction outcomes. Second, studies processing the retinal images with vessel segmentation algorithms to predict CVD risk were excluded in this review for the purpose of keeping the predictors homogeneous to facilitate the interpretation and preventing us from overlooking the value of other retinal features. Nevertheless, retinal vasculature is one of the most significant retinal features in predicting CVD risk, and numerous studies have been conducted on this specific topic, therefore, a review focusing on this topic is worthwhile to be conducted separately. Third, a limited number of studies were eligible for the meta-analysis which increases the uncertainties of the outcome of pooled estimates.[67] Finally, as no quality assessment tools are currently available for AI-based diagnostic accuracy studies, we used the QUADAS-2 quality assessment tool of which the interpretation can be varied between studies.

Conclusion

In conclusion, this systematic review and meta-analysis qualitatively interpret the studies that use DL and retinal images to predict CVD risk-related outcomes. A wide range of the predicted outcomes was investigated but the evidence is scarce on the prediction of incident CVD longitudinally. Future studies are needed to validate and refine the algorithms, especially in large-scale longitudinal cohorts. In addition, prospective studies need to be conducted to prove the applicability of the technology in real-world practice.

Acknowledgments

Supported by Medical Research Future Fund (MRFF; MRFAI000035), NHMRC Investigator Grants (APP1175405 and 2010072) and High-level Talent Flexible Introduction Fund of Guangdong Provincial People's Hospital (KJ012019530). The sponsor or funding organization had no role in the design or conduct of this research.

Author Contributions: Study concept and design: M.G.H., W.Y.H., and Z.T.Z. Access to raw data: W.Y.H. and R.Y.C. Acquisition, analysis, or interpretation: W.Y.H., F.Y., and R.Y.C. Drafting of the manuscript: W.Y.H. Critical revision of the manuscript for important intellectual content: M.G.H., A.V., Z.T.Z., X.W.S., K.K., and E.W. Statistical analysis: W.Y.H. and X.W.S. Obtained funding: M.G.H. Administrative, technical, or material support: M.G.H. Study supervision: M.G.H., A.V., Z.T.Z., X.H.Y., X.W.S., K.K., and E.W.

Disclosure: **W. Hu**, None; **F.S.L. Yii**, None; **R. Chen**, None; **X. Zhang**, None; **X. Shang**, None; **K. Kiburg**, None; **E. Woods**, None; **A. Vingrys**, None; **L. Zhang**, None; **Z. Zhu**, None; **M. He**, None

References

1. Stampfer MJ, Hu FB, Manson JE, et al. Primary prevention of coronary heart disease in women through diet and lifestyle. *N Engl J Med*. 2000;343(1):16–22.
2. Chiuve SE, McCullough ML, Sacks FM, Rimm EB. Healthy lifestyle factors in the primary prevention of coronary heart disease among men: benefits among users and nonusers of lipid-lowering and antihypertensive medications. *Circulation*. 2006;114(2):160–167.
3. Manuel DG, Lim J, Tanuseputro P, et al. Revisiting Rose: strategies for reducing coronary heart disease. *BMJ*. 2006;332(7542):659–662.
4. Heart Foundation. Cardiovascular disease (CVD) risk assessment and management. Available at: https://www.heartfoundation.org.au/bundles/for-professionals/cvd-risk-assessment-and-management. Accessed November 16, 2022.
5. D'Agostino RB, Sr., Vasan RS, Pencina MJ, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation*. 2008;117(6):743–753.

6. Group Sw, Collaboration ESCCr. SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. *Eur Heart J*. 2021;42(25):2439–2454.
7. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ*. 2017;357:j2099.
8. Persell SD, Dunne AP, Lloyd-Jones DM, Baker DW. Electronic health record-based cardiac risk assessment and identification of unmet preventive needs. *Med Care*. 2009;47(4):418–424.
9. Flammer J, Konieczka K, Bruno RM, et al. The eye and the heart. *Eur Heart J*. 2013;34(17):1270–1278.
10. Huang L, Aris IM, Teo LLY, et al. Exploring associations between cardiac structure and retinal vascular geometry. *J Am Heart Assoc*. 2020;9(7):e014654.
11. McGeechan K, Liew G, Macaskill P, et al. Meta-analysis: retinal vessel caliber and risk for coronary heart disease. *Ann Intern Med*. 2009;151(6):404–413.
12. Seidelmann SB, Claggett B, Bravo PE, et al. Retinal vessel calibers in predicting long-term cardiovascular outcomes: the atherosclerosis risk in communities study. *Circulation*. 2016;134(18):1328–1338.
13. Arnould L, Binquet C, Guenancia C, et al. Association between the retinal vascular network with Singapore "I" Vessel Assessment (SIVA) software, cardiovascular history and risk factors in the elderly: the montrachet study, population-based study. *PLoS One*. 2018;13(4):e0194694.
14. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436–444.
15. Raumviboonsuk P, Krause J, Chotcomwongse P, et al. Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. *NPJ Digit Med*. 2019;2:25.
16. Wong DYL, Lam MC, Ran A, Cheung CY. Artificial intelligence in retinal imaging for cardiovascular disease prediction: current trends and future directions. *Curr Opin Ophthalmol*. 2022;33(5):440–446.
17. Arnould L, Meriaudeau F, Guenancia C, et al. Using artificial intelligence to analyse the retinal vascular network: the future of cardiovascular risk assessment based on oculomics? *A Narrative Review. Ophthalmol Ther*. 2023;12(2):657–674.
18. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71.
19. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155(8):529–536.
20. Barriada RG, Simó-Servat O, Planas A, et al. Deep learning of retinal imaging: a useful tool for coronary artery calcium score prediction in diabetic patients. *Appl Sci (Switzerland)*. 2022;12(3):1401.
21. Son J, Shin JY, Chun EJ, et al. Predicting high coronary artery calcium score from retinal fundus images with deep learning algorithms. *Transl Vis Sci Technol*. 2020;9(6):28.
22. Rim TH, Lee CJ, Tham YC, et al. Deep-learning-based cardiovascular risk stratification using coronary artery calcium scores predicted from retinal photographs. *Lancet Digit Health*. 2021;3(5):e306–e316.
23. Islam MT, Al-Absi HRH, Ruagh EA, Alam T. DiaNet: a deep learning based architecture to diagnose diabetes using retinal images only. *IEEE Access*. 2021;9:15686–15695.
24. Diaz-Pinto A, Ravikumar N, Attar R, et al. Predicting myocardial infarction through retinal scans and minimal personal information. *Nat Mach Intell*. 2022;4(1):55–61.
25. Nusinovici S, Rim TH, Yu M, et al. Retinal photograph-based deep learning predicts biological age, and stratifies morbidity and mortality risk. *Age Ageing*. 2022;51(4):afac065.
26. Poplin R, Varadarajan AV, Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng*. 2018;2(3):158–164.
27. Nagasato D, Tabuchi H, Masumoto H, et al. Prediction of age and brachial-ankle pulse-wave velocity using ultra-wide-field pseudo-color images by deep learning. *Sci Rep*. 2020;10(1):19369.
28. Zhang L, Yuan M, An Z, et al. Prediction of hypertension, hyperglycemia and dyslipidemia from retinal fundus photographs via deep learning: a cross-sectional study of chronic diseases in central China. *PLoS One*. 2020;15(5):e0233166.
29. Mueller S, Wintergerst MWM, Falahat P, et al. Multiple instance learning detects peripheral arterial disease from high-resolution color fundus photography. *Sci Rep*. 2022;12(1):1389.
30. Vaghefi E, Yang S, Hill S, et al. Detection of smoking status from retinal images; a convolutional neural network study. *Sci Rep*. 2019;9(1):7180.
31. Kim YD, Noh KJ, Byun SJ, et al. Effects of hypertension, diabetes, and smoking on age and sex prediction from retinal fundus images. *Sci Rep*. 2020;10(1):4623.

32. Gerrits N, Elen B, Craenendonck TV, et al. Age and sex affect deep learning prediction of cardiometabolic risk factors from retinal images. *Sci Rep*. 2020;10:9432.
33. Heslinga FG, Pluim JPW, Houben AJHM, et al. Direct classification of type 2 diabetes from retinal fundus images in a population-based sample from the Maastricht study. In *Medical Imaging 2020: Computer-Aided Diagnosis*. 2020; v. 11314: pp. 383–388. SPIE.
34. Zhu ZT, Shi DL, Guankai P, et al. Retinal age gap as a predictive biomarker for mortality risk. *Br J Ophthalmol*. 2023;107.4:547–554.
35. Ma Y, Xiong J, Zhu Y, et al. Deep learning algorithm using fundus photographs for 10-year risk assessment of ischemic cardiovascular diseases in China. *Science Bull*. 2022;67:17–20.
36. Kang EYC, Hsieh YT, Li CH, et al. Deep learning-based detection of early renal function impairment using retinal fundus images: model development and validation. *JMIR Medical Informatics*. 2020;8(11):e23472.
37. Sabanayagam C, Xu D, Ting DSW, et al. A deep learning algorithm to detect chronic kidney disease from retinal photographs in community-based populations. *The Lancet Digital Health*. 2020;2(6):e295–e302.
38. Zhang K, Liu X, Xu J, et al. Deep-learning models for the detection and incidence prediction of chronic kidney disease and type 2 diabetes from retinal fundus images. *Nat Biomed Eng*. 2021;5(6):533–545.
39. Al-Absi HRH, Islam MT, Refaee MA, et al. Cardiovascular disease diagnosis from DXA scan and retinal images using deep learning. *Sensors*. 2022;22(12):4310.
40. Yun JS, Kim J, Jung SH, Cha SA, Ko SH, Ahn YB, Won HH, Sohn KA, Kim D. A deep learning model for screening type 2 diabetes from retinal photographs. Nutrition, metabolism, and cardiovascular diseases. *Nutr Metab Cardiovasc Dis*. 2022;32(5):1218–1226.
41. Syed MG, Doney A, George G, et al. Are cardiovascular risk scores from genome and retinal image complementary? A deep learning investigation in a diabetic cohort. Ophthalmic Medical Image Analysis. OMIA 2021. Lecture Notes in Computer Science(), vol 12970. pp. 109–118.
42. Chang J, Shin JY, Ko T, et al. Association of deep learning-based fundus age difference with carotid atherosclerosis and mortality. *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, San Diego: CA, USA, 2019, pp. 1179–1181.
43. Chang J, Ko A, Park SM, et al. Association of cardiovascular mortality and deep learning-funduscopic atherosclerosis score derived from retinal fundus images. *Am J Ophthalmol*. 2020; 217:121–130.
44. Khan NC, Perera C, Dow ER, et al. Predicting systemic health features from retinal fundus images using transfer-learning-based artificial intelligence models. *Diagnostics (Basel)*. 2022;12(7): 1714.
45. Coronado I, Abdelkhaleq R, Yan J, et al. Towards stroke biomarkers on fundus retinal imaging: a comparison between vasculature embeddings and general purpose convolutional neural networks. *Annu Int Conf IEEE Eng Med Biol Soc*. 2021:3873–3876.
46. Reinikainen J, Laatikainen T, Karvanen J, Tolonen H. Lifetime cumulative risk factors predict cardiovascular disease mortality in a 50-year follow-up study in Finland. *Int J Epidemiol*. 2015;44(1):108–116.
47. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology*. 2018;286(3):800–809.
48. Sounderajah V, Ashrafian H, Golub RM, et al. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open*. 2021;11(6):e047709.
49. Schmiegelow MD, Hedlin H, Mackey RH, et al. Race and ethnicity, obesity, metabolic health, and risk of cardiovascular disease in postmenopausal women. *J Am Heart Assoc*. 2015;4(5): e001695.
50. Hozawa A, Folsom AR, Sharrett AR, Chambless LE. Absolute and attributable risks of cardiovascular disease incidence in relation to optimal and borderline risk factors: comparison of African American with white subjects–Atherosclerosis Risk in Communities Study. *Arch Intern Med*. 2007;167(6):573–579.
51. Anand SS, Yusuf S, Jacobs R, et al. Risk factors, atherosclerosis, and cardiovascular disease among Aboriginal people in Canada: the study of health assessment and risk evaluation in aboriginal peoples (SHARE-AP). *Lancet*. 2001;358(9288):1147–1153.
52. Gardiner FW, Rallah-Baker K, Dos Santos A, et al. Indigenous Australians have a greater prevalence of heart, stroke, and vascular disease, are younger at death, with higher hospitalisation and more aeromedical retrievals from remote regions. *EClinicalMedicine*. 2021;42:101181.

53. Guan H, Liu M. Domain adaptation for medical image analysis: a survey. *IEEE Trans Biomed Eng*. 2022;69(3):1173–1185.
54. Muftuoglu IK, Gaber R, Bartsch DU, et al. Comparison of conventional color fundus photography and multicolor imaging in choroidal or retinal lesions. *Graefes Arch Clin Exp Ophthalmol*. 2018;256(4):643–649.
55. Fenner BJ, Wong RLM, Lam WC, et al. Advances in retinal imaging and applications in diabetic retinopathy screening: a review. *Ophthalmol Ther*. 2018;7(2):333–346.
56. Das S, Kuht HJ, De Silva I, et al. Feasibility and clinical utility of handheld fundus cameras for retinal imaging. *Eye (Lond)*. 2022;37(2):274–279.
57. Ruamviboonsuk P, Tiwari R, Sayres R, et al. Real-time diabetic retinopathy screening by deep learning in a multisite national screening programme: a prospective interventional cohort study. *Lancet Digit Health*. 2022;4(4):e235–e244.
58. Ipp E, Liljenquist D, Bode B, et al. Pivotal evaluation of an artificial intelligence system for autonomous detection of referrable and vision-threatening diabetic retinopathy. *JAMA Netw Open*. 2021;4(11):e2134254.
59. Abramoff MD, Lavin PT, Birch M, et al. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med*. 2018;1:39.
60. Kokol P, Kokol M, Zagoranski S. Machine learning on small size samples: a synthetic knowledge synthesis. *Sci Prog*. 2022;105(1):3685042110 29777.
61. Chia YC, Gray SY, Ching SM, et al. Validation of the Framingham general cardiovascular risk score in a multiethnic Asian population: a retrospective cohort study. *BMJ Open*. 2015;5(5): e007324.
62. Qureshi WT, Michos ED, Flueckiger P, et al. Impact of replacing the pooled cohort equation with other cardiovascular disease risk scores on atherosclerotic cardiovascular disease risk assessment (from the Multi-Ethnic Study of Atherosclerosis [MESA]). *Am J Cardiol*. 2016;118(5):691–696.
63. Rabanal KS, Meyer HE, Pylypchuk R, et al. Performance of a Framingham cardiovascular risk model among Indians and Europeans in New Zealand and the role of body mass index and social deprivation. *Open Heart*. 2018;5(2): e000821.
64. Siontis GC, Tzoulaki I, Siontis KC, Ioannidis JP. Comparisons of established risk prediction models for cardiovascular disease: systematic review. *BMJ*. 2012;344:e3318.
65. Harish KB, Price WN, Aphinyanaphongs Y. Open-source clinical machine learning models: critical appraisal of feasibility, advantages, and challenges. *JMIR Form Res*. 2022;6(4):e33970.
66. Sing CF, Stengard JH, Kardia SL. Genes, environment, and cardiovascular disease. *Arterioscler Thromb Vasc Biol*. 2003;23(7):1190–1196.
67. Goh JX, Hall JA, Rosenthal R. Mini meta-analysis of your own studies: some arguments on why and a primer on how. *Social and Personality Psychology Compass*. 2016;10(10):535–549.