



THE AGA KHAN UNIVERSITY

eCommons@AKU

---

Haematology and Oncology, East Africa

Medical College, East Africa

---

4-2024

## Prediction of cardiovascular risk factors from retinal fundus photographs: Validation of a deep learning algorithm in a prospective non-interventional study in Kenya

Tom White  
*AstraZeneca, UK*

Viknesh Selvarajah  
*AstraZeneca, UK*

Fredrik Wolfhagen  
*AstraZeneca, UK*

Nils Svängård  
*AstraZeneca, UK*

Gayathri Mohankumar  
*AstraZeneca, USA*

*See next page for additional authors*

Follow this and additional works at: [https://ecommons.aku.edu/eastafrica\\_fhs\\_mc\\_haematol\\_oncol](https://ecommons.aku.edu/eastafrica_fhs_mc_haematol_oncol)



Part of the [Cardiology Commons](#)

---

### Recommended Citation

White, T., Selvarajah, V., Wolfhagen, F., Svängård, N., Mohankumar, G., Fenici, P., Rough, K., Onyango, N., Saleh, M., Abayo, I. (2024). Prediction of cardiovascular risk factors from retinal fundus photographs: Validation of a deep learning algorithm in a prospective non-interventional study in Kenya. *Diabetes, Obesity and Metabolism*, 1-10.




Available at: [https://ecommons.aku.edu/eastafrica\\_fhs\\_mc\\_haematol\\_oncol/2](https://ecommons.aku.edu/eastafrica_fhs_mc_haematol_oncol/2)

---

**Authors**

Tom White, Viknesh Selvarajah, Fredrik Wolfhagen, Nils Svangård, Gayathri Mohankumar, Peter Fenici, Kathryn Rough, Nelson Onyango, Mansoor Saleh, and Innocent Abayo

# Prediction of cardiovascular risk factors from retinal fundus photographs: Validation of a deep learning algorithm in a prospective non-interventional study in Kenya

Tom White PhD<sup>1</sup> | Viknesh Selvarajah PhD<sup>2</sup>  | Fredrik Wolfhagen-Sand PhD<sup>2</sup>  | Nils Svängård MSc<sup>3</sup> | Gayathri Mohankumar MS<sup>4</sup> | Peter Fenici MD<sup>5,6,7</sup> | Kathryn Rough ScD<sup>8</sup> | Nelson Onyango PhD<sup>8</sup> | Kendall Lyons MPH<sup>8</sup> | Christina Mack PhD<sup>8</sup> | Videlis Nduba PhD<sup>9</sup> | Mansoor Noorali Saleh MD<sup>10</sup> | Innocent Abayo BSc<sup>10</sup> | Afrah Siddiqui MBBS<sup>11</sup> | Malgorzata Majdanska-Strzalka MPharm<sup>12</sup> | Katarzyna Kaszubska PhD<sup>12</sup> | Tove Hegelund-Myrback PhD<sup>13</sup> | Russell Esterline PhD<sup>14</sup> | Antonio Manzur BS<sup>2</sup> | Victoria E. R. Parker PhD<sup>2</sup> 

<sup>1</sup>Data Science and Advanced Analytics, Data Science & Artificial Intelligence, R&D, AstraZeneca, Cambridge, UK

<sup>2</sup>Research and Early Development, Cardiovascular, Renal and Metabolism (CVRM), BioPharmaceuticals R&D, AstraZeneca, Cambridge, UK

<sup>3</sup>Data Science and Advanced Analytics, Data Science & Artificial Intelligence, R&D, AstraZeneca, Gothenburg, Sweden

<sup>4</sup>Centre for Artificial Intelligence, Data Science & Artificial Intelligence, R&D, AstraZeneca, Gaithersburg, Maryland, USA

<sup>5</sup>School of Medicine and Surgery, Catholic University, Rome, Italy

<sup>6</sup>Biomagnetism and Clinical Physiology International Center (BACPIC), Rome, Italy

<sup>7</sup>AstraZeneca, Medical Affairs, BioPharmaceuticals, AstraZeneca, Milan, Italy

<sup>8</sup>IQVIA, Durham, North Carolina, USA

<sup>9</sup>Kenya Medical Research Institute, Nairobi, Kenya

<sup>10</sup>Clinical Research Unit, Aga Khan University Hospital, Nairobi, Kenya

<sup>11</sup>BioPharmaceuticals Medical, AstraZeneca, Cambridge, UK

<sup>12</sup>CVRM Clinical Operations, Biopharmaceuticals, R&D, AstraZeneca, Warsaw, Poland

<sup>13</sup>Global Portfolio & Project Management, Early CVRM&NS, R&D, AstraZeneca, Gothenburg, Sweden

<sup>14</sup>Research and Late Development, Cardiovascular, Renal and Metabolism (CVRM), BioPharmaceuticals R&D, AstraZeneca, Gaithersburg, USA

## Correspondence

Victoria E. R. Parker, Research and Early Development, Cardiovascular, Renal and Metabolism (CVRM), BioPharmaceuticals R&D, AstraZeneca, Cambridge, UK.  
Email: [victoria.parker@astrazeneca.com](mailto:victoria.parker@astrazeneca.com)

**Funding information**  
AstraZeneca

## Abstract

**Aim:** Hypertension and diabetes mellitus (DM) are major causes of morbidity and mortality, with growing burdens in low-income countries where they are underdiagnosed and undertreated. Advances in machine learning may provide opportunities to enhance diagnostics in settings with limited medical infrastructure.

Tom White and Viknesh Selvarajah contributed equally.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 AstraZeneca. *Diabetes, Obesity and Metabolism* published by John Wiley & Sons Ltd.

**Materials and Methods:** A non-interventional study was conducted to develop and validate a machine learning algorithm to estimate cardiovascular clinical and laboratory parameters. At two sites in Kenya, digital retinal fundus photographs were collected alongside blood pressure (BP), laboratory measures and medical history. The performance of machine learning models, originally trained using data from the UK Biobank, were evaluated for their ability to estimate BP, glycated haemoglobin, estimated glomerular filtration rate and diagnoses from fundus images.

**Results:** In total, 301 participants were enrolled. Compared with the UK Biobank population used for algorithm development, participants from Kenya were younger and would probably report Black/African ethnicity, with a higher body mass index and prevalence of DM and hypertension. The mean absolute error was comparable or slightly greater for systolic BP, diastolic BP, glycated haemoglobin and estimated glomerular filtration rate. The model trained to identify DM had an area under the receiver operating curve of 0.762 (0.818 in the UK Biobank) and the hypertension model had an area under the receiver operating curve of 0.765 (0.738 in the UK Biobank).

**Conclusions:** In a Kenyan population, machine learning models estimated cardiovascular parameters with comparable or slightly lower accuracy than in the population where they were trained, suggesting model recalibration may be appropriate. This study represents an incremental step toward leveraging machine learning to make early cardiovascular screening more accessible, particularly in resource-limited settings.

#### KEYWORDS

cardiovascular screening, diabetes mellitus, hypertension, machine learning, predictive modelling, retinal fundus photographs, UK Biobank

## 1 | INTRODUCTION

Hypertension and diabetes mellitus (DM) are major causes of morbidity and mortality; both are within the top 10 causes of death worldwide.<sup>1,2</sup> In low-resource settings, societal shifts and urbanization have driven a surge in morbidity and mortality from cardiovascular disease.<sup>2,3</sup> Since 2000, DM has increased by 70% globally and the prevalence of hypertension has shifted from wealthy countries to low- and middle-income countries.<sup>1</sup> While numerous lifestyle and pharmaceutical interventions are available to treat hypertension and DM, diagnosis often requires repeated, office-based measurements.<sup>4</sup> High-touch medical requirements present a challenge in low-income countries; up to two-thirds of adults are undiagnosed or untreated for DM and hypertension.<sup>5</sup> This problem is compounded by a lack of health care providers in these parts of the world.<sup>6</sup>

Advances in machine learning and decreasing imaging costs may provide opportunities to improve access to cost-effective diagnostics in settings with limited health care infrastructure through artificial intelligence-based screening.<sup>7,8</sup> Machine learning algorithms have been trained to diagnose diabetic retinopathy from digital retinal images,<sup>9–12</sup> to estimate blood pressure (BP)<sup>13–16</sup> and glycated

haemoglobin (HbA1c),<sup>13,14,17</sup> and to diagnose DM<sup>15,18,19</sup> and hypertension.<sup>16,20</sup> Further machine learning-based technology development could deliver rapid, non-invasive diagnosis that is deployable in a wide variety of settings. Improved diagnosis could facilitate earlier intervention, leading to tangible health improvements in resource-limited settings. Such an intervention could be particularly beneficial in African populations, where undiagnosed hypertension and DM are particularly severe.<sup>21–23</sup>

Studies have shown the feasibility of training machine learning algorithms to estimate a variety of cardiovascular and laboratory measures based on retinal images.<sup>13,24</sup> The model performance varies between studies and much of the research is done retrospectively, with little emphasis on evaluating algorithms in the low-resource settings that might benefit most from these types of diagnostic tools.<sup>13,24</sup>

The goal of this study was to show feasibility in a Kenyan population and exploratory objectives were to train a deep learning algorithm to estimate clinical and laboratory parameters, including hypertension and DM from retinal fundus photographs and to show the feasibility of validating the algorithm performance in an African population.

## 2 | MATERIALS AND METHODS

### 2.1 | Study design overview

Retrospective data from the UK Biobank were used to train a series of machine learning algorithms to estimate clinical and laboratory parameters, which were prospectively validated in a non-interventional study in Kenya.

The UK Biobank is a large, population-based prospective study.<sup>25</sup> Between 2006 and 2010, it enrolled 500 000 volunteers aged 40-69 years old. Participants provided biological samples and detailed health information and consented to having their health outcomes prospectively followed up. Figure 1 displays a study design schematic.

After finalization of the machine learning algorithms, a non-interventional cohort study was conducted at two sites in Kenya with the objective of establishing the feasibility of data collection and prospectively evaluate the algorithms in an African population. The prospective validation part of the study was registered at [clinicaltrials.gov](https://clinicaltrials.gov) (NCT04814680) and was approved by both sites' local Independent Ethics Committees.

### 2.2 | Study population

For model development, all UK Biobank participants  $\geq 40$  years of age with retinal image data and sufficient clinical data were included in the training of each model, and retinal fundus image quality was algorithmically determined. No additional eligibility criteria were imposed.

In the prospective validation component of the study, adults aged  $>35$  years and over and at least 50% needed an HbA1c  $\geq 6.5\%$  or a diagnosis of DM. Inclusion and exclusion criteria are listed in Data S1 (page 8). Participants who had an eye condition known to preclude clear retinal imaging were excluded. During a routine clinical visit, participants who met eligibility criteria and provided written informed consent were enrolled. Referral health care facilities were utilized to recruit participants; in addition, participants were recruited with the help of local diabetic support groups. The study was conducted according to the principles of the Declaration of Helsinki and the

International Council for Harmonization Guidance for Good Clinical Practice.

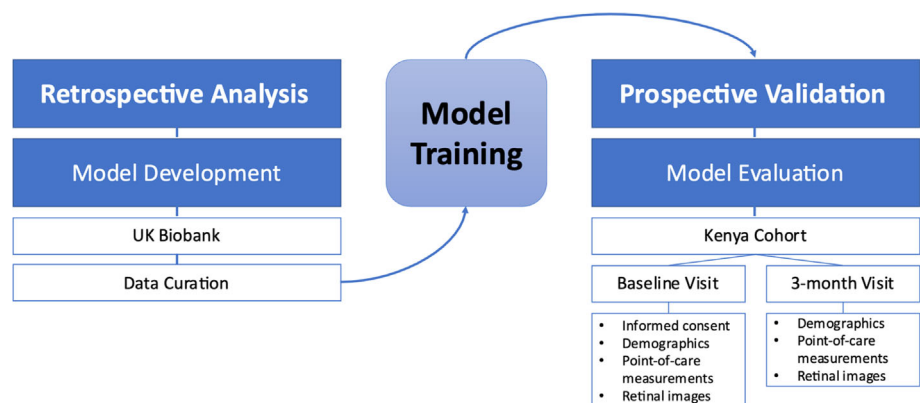
### 2.3 | Data collection

In the UK Biobank, fundus photographs had a  $45^\circ$  primary field of view and were taken with a TOPCON 3D OCT 1000 Mk2 device. Images were graded for quality based on mean pixel intensity, which resulted in the inclusion of 135 359/175 831 images from 70 984 UK Biobank participants. Implausible outliers were removed for numerical data, including lab values.

For the prospective Kenyan validation cohort, study eligibility was determined at a baseline visit and retinal fundus images, clinical measurements, demographics, medical history and samples for laboratory testing were collected. During visits, BP was measured three times per participant (at  $60 \pm 15$  min before,  $15 \pm 5$  min before/after, and  $60 \pm 15$  min after retinal imaging). All BP measurements occurred before blood sampling. For HbA1c, both point-of-care and laboratory testing were conducted; laboratory testing results are reported in this manuscript. A Canon CR-2AF fundus camera equipped with a Canon EOS camera back (Canon RX Capture software) was used for retinal image acquisition. Retinal images from the prospective validation cohort were manually reviewed and assigned quality ratings by retinal imaging experts (Merit©).

### 2.4 | Model training

Deep learning algorithms were developed exclusively from UK Biobank data. The data were partitioned into three subsets on a per-participant-per-time point basis: a training set (86.7%), a validation set (8.5%) and a held-out test set (4.6%). The unconventional bias in the data split is because of the imbalance of healthy to unhealthy patients in the UK Biobank dataset, resulting in fewer unhealthy patients. The training and validation sets were used to iteratively update and tune the model. The held-out test set was reserved for evaluating the final model and was otherwise unused during model training and tuning. No data from the prospective validation cohort were used to train the models.



**FIGURE 1** Illustration of study design.

Separate machine learning models were trained to estimate current systolic BP (SBP), diastolic BP (DBP), HbA1c, creatinine, cystatin C, DM and hypertension diagnoses based on retinal images. Model-produced creatinine and cystatin C estimations were used to calculate the estimated glomerular filtration rate (eGFR).<sup>26</sup>

The model training procedures closely followed those described by Poplin et al.<sup>14</sup> The main objective of the prediction task for the model is to give an input retinal fundus image of sufficient quality and resize it to 224 × 224 pixels to predict the disease status or interested biomarker. For each prediction task, we used a deep learning architecture, Inception v3.<sup>27</sup> The base model was the previously published Inception V3 model architecture, pretrained on ImageNet-1K from torchvision.<sup>28</sup> The final fully connected and auxiliary logit layers were replaced with a linear layer of outputs equal to the number of targets predicted by the model. Binary cross entropy loss and mean squared error was used for optimizing the models for predicting disease status and for biomarker prediction respectively. We tuned the model parameters with various optimizers and learning rates using the validation set. The Adam optimizer with a learning rate of 3e-4 performed the best on the validation dataset and were used to train the final models for each task. The training data set was presented to the model during training with a batch size of 16 for 50 epochs. The model checkpoint with a stable validation loss across 10 epochs was used as early stopping criteria, and the checkpoint was used for internal and external evaluation. All the experiments were run on a cloud computed with 1 GPU core with a memory 16GB. All trained models took a single retinal image as input and generated predictions on a per-image basis.

Model training was completed using Python v3.7 and Pytorch v1.3.

## 2.5 | Model evaluation

For model evaluation, predictions were aggregated on a per-participant/per-time point basis. Retinal fundus photographs deemed inadequate (algorithmically for the UK Biobank; manual review for the prospective validation cohort) were excluded from training and evaluation. If multiple images were available for a participant at a given time point, participant-level predictions were derived by taking the mean of the per-image predictions. The model evaluation was performed on the held-out test and a prospective validation cohort with a balanced disease distribution for robust performance evaluation.

For continuous predictions, root mean squared error, mean absolute error (MAE) and mean bias error (MBE) were calculated. MAE summarizes the magnitude of the average model error compared with the measured values, while MBE quantifies the magnitude and direction of the bias in errors by subtracting predicted values from the observed values. The accuracies of SBP, DBP and HbA1c were assessed by evaluating whether the algorithm output was within clinically relevant windows of error. These windows were defined a priori, based on expert clinician input.

For binary prediction tasks (i.e. diagnoses of hypertension and DM), the area under the receiving operating curve (AUROC),

sensitivity, specificity, positive predictive value (PPV) and negative predictive value were calculated. The model-predicted presence of the condition corresponded to a predicted probability exceeding a specified threshold. Otherwise, the model was considered to have predicted the absence of the condition. Thresholds for binary predictions were selected based on results in the UK Biobank training and validation data before prospective data collection.

The evaluation results are presented for both the UK Biobank test set data and the prospective validation cohort. The sample size determination was empirical, and no formal statistical comparisons were made, nor were any hypothesis tests conducted. All evaluations were conducted using Python v3.7.

## 3 | RESULTS

### 3.1 | Participants

#### 3.1.1 | UK Biobank

In the UK Biobank data, 70 984 unique participants with 135 522 retinal fundus photographs contributed to analyses. The mean age of participants was 58 years old, and more women (53%) than men were included. The mean body mass index was 27.3 kg/m<sup>2</sup>, and >90% of participants self-identified as British, Irish, or of other White background. In the test set, 3.2% of participants had a documented DM diagnosis, and 11.8% had a hypertensive disease diagnosis at baseline. Baseline characteristics of the UK Biobank training and test sets are presented in Table 1.

#### 3.1.2 | Kenyan cohort

In the prospective Kenyan validation cohort, the first participant was enrolled on 8 November 2021 and follow-up was completed by 10 February 2022. Of the 317 screened, 301 participants from two clinical trial sites in Kenya were found eligible and consented. Nine participants discontinued or withdrew from the study because of participant choice (n = 2), SAE (n = 1) or other reasons (n = 6). Participants had a mean age of 51 years, and more men (54.5%) than women enrolled (Table 1). The mean body mass index was 28.12 kg/m<sup>2</sup>. Approximately 99% of the study population reported their race as Black or African, and <1% as Asian. Nearly half (n = 147) of participants had either a baseline HbA1c measurement of ≥6.5% or a known diagnosis of DM, and 44% of participants had a hypertension diagnosis.

### 3.2 | Outcome measures

#### 3.2.1 | Data completeness

The primary objective of the study was to assess the successful use of methodology for retinal image acquisition, BP, HbA1c and eGFR

**TABLE 1** Descriptive and baseline medical characteristics of participants from the UK Biobank and prospective Kenyan validation cohorts.

Characteristic	UK Biobank		Kenyan prospective validation cohort (N = 301)
	Training set (N = 56 705)	Held-out test set (N = 4076)	
Age, years			
Mean (SD)	57.7 (8.2)	57.6 (8.3)	51.1 (9.5)
Median (Q1, Q3)	59.0 (51.0, 64.0)	59.0 (51.0, 64.0)	51.0 (43.0, 58.0)
Minimum, maximum	40.0, 79.0	40.0, 75.0	35.0, 85.0
Sex, n (%)			
Male	26 745 (47.0)	1889 (46.3)	164 (54.5)
Female	29 960 (53.0)	2187 (53.7)	137 (45.5)
BMI, kg/m <sup>2</sup>			
Mean (SD)	27.26 (4.69)	27.28 (4.69)	28.12 (5.92)
Median (Q1, Q3)	26.61 (24.02, 29.70)	26.63 (24.09, 29.72)	27.20 (23.8, 31.3)
Minimum, maximum	12.65, 65.01	15.11, 59.38	16.0, 52.2
Race and ethnicity, n (%)			
White <sup>a</sup>	52 475 (92.5)	3767 (92.4)	0 (0)
Asian <sup>b</sup>	1538 (2.7)	116 (2.8)	2 (0.7)
Black or African <sup>c</sup>	1312 (2.3)	93 (2.3)	299 (99.3)
Other	1380 (2.4)	91 (2.2)	0 (0.0)
SBP at baseline, <sup>d</sup> mmHg			
Mean (SD)	137.5 (18.4)	137.1 (18.4)	131.6 (18.4)
Median (Q1, Q3)	136.0 (124.5, 149.0)	135.5 (124.0, 148.5)	130.0 (118.0, 140.0)
Minimum, maximum	76.5, 239.0	77.5, 245.0	100.0, 220.0
DBP at baseline, <sup>d</sup> mmHg			
Mean (SD)	81.5 (9.9)	81.1 (9.9)	82.5 (11.4)
Median (Q1, Q3)	81.0 (74.5, 88.0)	81.0 (74.0, 87.5)	82.3 (74.0, 89.0)
Minimum, maximum	46.5, 134.0	49.5, 122.5	57.0, 139.0
HbA1c at baseline, %			
Mean (SD)	5.5 (0.6)	5.4 (0.6)	7.6 (2.7)
Median (Q1, Q3)	5.5 (5.1, 5.6)	5.4 (5.1, 5.6)	6.2 (5.6, 9.2)
Minimum, maximum	3.6, 9.2	3.7, 13.4	4.0, 17.6
eGFR, creatinine at baseline; ml/min/1.73 m <sup>2</sup>			
Mean (SD)	89.3 (13.5)	87.8 (16.0)	96.3 (16.8)
Median (Q1, Q3)	91.0 (81.2, 98.6)	88.4 (76.9, 100.2)	100.5 (85.8, 108.5)
Minimum, maximum	17.5, 132.3	14.7, 137.1	28.1, 128.3
DM at baseline, n (%)	1745 (3.1%)	129 (3.2%)	138 (45.8%)
Hypertension at baseline, n (%)	6332 (11.2%)	465 (11.4%)	132 (43.9%)

Abbreviations: BMI, body mass index; DBP, diastolic blood pressure; DM, diabetes mellitus, eGFR, estimated glomerular filtration rate; HbA1c, glycated haemoglobin; Q1, 25th percentile; Q3, 75th percentile; SBP, systolic blood pressure; SD, standard deviation.

<sup>a</sup>For the UK Biobank, this is the sum of participants who indicated 'British', 'Irish', or 'other White background'.

<sup>b</sup>For the UK Biobank, this is the sum of participants who indicated 'Indian', 'Chinese', 'Pakistani', 'Bangladeshi', or 'any other Asian background'.

<sup>c</sup>For the UK Biobank, this is the sum of the participants who indicated 'African', 'Caribbean', 'Black or Black British', or 'Any other Black' background'.

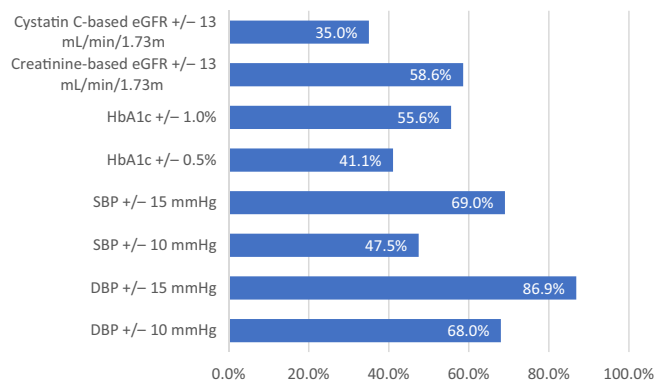
<sup>d</sup>For Kenyan prospective validation cohort participants, reported blood pressures are the average of taken 60 ± 15 min before retinal imaging, 15 ± 5 min before or after retinal imaging, and 60 ± 15 min after retinal imaging.

measurement and to show the feasibility of validating the algorithm's performance in an African population. Table S2 reports the feasibility results for the prospective Kenyan validation cohort. All 301 participants completed the baseline visit, and 292 completed the follow-up

visit (97%). All study procedures had a high level of completeness, with all protocol-mandated measurements occurring in >97% of participants at each visit. For >98% of participants, complete and interpretable images of both eyes of sufficient quality were available for each visit.

Parameter	MBE (95% CI)	MAE	RMSE
<b>HbA1c, %</b>			
UK Biobank test set	0.00 (−1.09, 1.09)	0.34	0.56
Kenyan validation cohort, baseline visit	−1.69 (−6.79, 3.40)	1.96	3.10
Kenyan validation cohort, follow-up visit	−1.78 (−6.99, 3.43)	1.94	3.19
<b>SBP, mmHg</b>			
UK Biobank test set	−0.63 (−29.35, 28.10)	11.41	14.67
Kenyan validation cohort, baseline visit	0.83 (−30.86, 32.51)	12.32	16.16
Kenyan validation cohort, follow-up visit	3.01 (−27.70, 33.82)	12.38	15.95
<b>DBP, mmHg</b>			
UK Biobank test set	−0.33 (−16.58, 15.91)	6.54	8.30
Kenyan validation cohort, baseline visit	3.06 (−16.00, 22.12)	7.98	10.18
Kenyan validation cohort, follow-up visit	6.10 (−14.84, 27.04)	9.99	12.29
<b>eGFR, creatinine; ml/min/1.73 m<sup>2</sup></b>			
UK Biobank test set	−2.09 (−33.41, 29.22)	12.72	16.11
Kenyan validation cohort, baseline visit	2.05 (−34.47, 38.57)	14.33	18.72
Kenyan validation cohort, follow-up visit	4.36 (−29.54, 38.25)	13.73	17.80
<b>eGFR, cystatin C; ml/min/1.73 m<sup>2</sup></b>			
UK Biobank test set	0.21 (−33.04, 33.45)	13.42	16.96
Kenyan validation cohort, baseline visit	19.34 (−17.35, 56.03)	22.36	26.89
Kenyan validation cohort, follow-up visit	21.46 (−18.92, 61.84)	24.98	29.72

Abbreviations: CI, confidence interval; DBP, diastolic blood pressure; eGFR, estimated glomerular filtration rate; HbA1c, glycated haemoglobin; MAE, mean absolute error; MBE, mean bias error; mmHg, millimetres mercury; RMSE, root mean square error; SBP, systolic blood pressure.



**FIGURE 2** Proportion of machine learning predictions of continuous measurements that fell within pre-specified tolerance ranges in the prospective Kenyan validation cohort. DBP, diastolic blood pressure; eGFR, estimated glomerular filtration rate; HbA1c, glycated haemoglobin; mmHg, millimetres mercury; SBP, systolic blood pressure.

### 3.2.2 | Machine learning predictions

Machine learning predictions of continuous measures, threshold values and binary endpoints were exploratory outcomes. Metrics of model performance for estimating continuous clinical and laboratory parameters are presented in Table 2, while Figure 2 summarizes the

**TABLE 2** Model performance results for estimating continuous clinical and laboratory parameters from retinal fundus photographs in the UK Biobank and prospective Kenyan validation cohorts.

proportion of machine learning algorithms' predictions of continuous measurements that fell within pre-specified tolerance ranges of measured values. The evaluation of models predicting binary endpoints is presented in Table 3. In the prospective validation cohort, analyses were also performed to assess whether the model-predicted values of HbA1c above a given threshold could be used to indirectly identify DM and whether model-predicted values of DBP or SBP could be used to indirectly identify hypertension.

#### *Blood pressure and hypertension diagnosis*

In the UK Biobank training dataset, the measured mean SBP was 137.5 mmHg (SD 18.4) and DBP was 81.5 mmHg (SD 9.9), and in the Kenyan, the cohort mean SBP was 131.6 mmHg (SD 18.4) and DBP was 82.5 mmHg (SD 11.4) at baseline. To evaluate if the temporal relationship between retinal image acquisition and BP measurement enhanced the accuracy of predictions, BPs were performed three times within a 60-min window before and after the retinal image. Using the mean of the three measurements, the MAE for estimation of SBP was similar between the prospective validation cohort at baseline (12.32 mmHg) and the UK Biobank (11.41 mmHg), although the MAE for DBP was marginally higher in the prospective validation cohort (7.98 vs. 6.54). In the validation cohort, minimal bias was observed in SBP (MBE, baseline: 0.83 mmHg), and a small positive bias was observed for DBP (MBE, baseline: 3.06 mmHg). At the baseline visit, for both SBP and DBP, the root mean squared error and



**TABLE 3** Model performance results for estimating binary clinical parameters from retinal fundus photographs in the UK Biobank and prospective Kenyan validation cohorts.

Parameter	AUROC	Sensitivity <sup>a</sup>	Specificity <sup>a</sup>	PPV <sup>a</sup>	NPV <sup>a</sup>
Agreement with diagnosis of DM in medical history					
Model-predicted DM					
UK Biobank test set	0.728	0.671	0.663	0.067	0.982
Kenyan validation cohort, baseline visit	0.762	0.801	0.534	0.592	0.761
Model-predicted HbA1c ≥6.5%					
Kenyan validation cohort, baseline visit	-	0.250	0.944	0.791	0.598
Model-predicted HbA1c ≥7.0%					
Kenyan validation cohort, baseline visit	-	0.110	1.000	1.000	0.571
Agreement with diagnosis of hypertension in medical history					
Model-predicted hypertension					
UK Biobank test set	0.687	0.686	0.580	0.187	0.929
Kenyan validation cohort, baseline visit	0.765	0.654	0.766	0.685	0.740
Model-predicted SBP ≥140 mmHg					
Kenyan validation cohort, baseline visit	-	0.477	0.838	0.697	0.673
Model-predicted DBP ≥80 mmHg					
Kenyan validation cohort, baseline visit	-	0.800	0.359	0.493	0.698

Abbreviations: AUROC, area under the receiver operating curve; DBP, diastolic blood pressure; DM, diabetes mellitus; eGFR, estimated glomerular filtration rate; HbA1c, glycated haemoglobin; NPV, negative predictive value; PPV, positive predictive value; SBP, systolic blood pressure.

<sup>a</sup>Values at or above the threshold were considered positive predictions for calculation of binary agreement metrics. For DM models, the threshold was a predicted probability  $\geq 0.142$ . For hypertension models, the threshold was a predicted probability  $\geq 0.047$ .

MAE were lower for the mean of three measures across time points compared with predictions of individual time points.

For DBP at baseline, 87% of model-produced estimates fell within 15 mmHg of the mean measured value, and 61% fell within 10 mmHg. For SBP, 69% of model predictions fell within the 15 mmHg tolerance limit and 49% fell within the 10 mmHg tolerance limit.

The model trained to identify patients with hypertension diagnoses had better discriminative abilities in the prospective validation cohort, with an AUROC of 0.77 versus 0.69. In the UK Biobank, using a threshold of  $\geq 0.047$  for hypertension resulted in a sensitivity of 0.69, a specificity of 0.58 and a PPV of 0.19. In the prospective validation cohort, the model had a sensitivity of 0.65, a specificity of 0.77 and a PPV of 0.69 at the same threshold. Using the DBP model with a threshold of 80 mmHg led to a hypertension diagnosis classifier with high sensitivity (0.80) but lower specificity (0.35). Using the SBP model with a threshold of 140 mmHg resulted in a lower sensitivity (0.48) and a higher specificity (0.84).

#### Glycated haemoglobin and diabetes diagnosis

The measured mean HbA1c was 5.5% (SD 0.6, range 3.6%-9.2%) in the UK Biobank and 7.6% (SD 2.7, range 4.0%-17.6%) in the Kenyan cohort at baseline. In the UK Biobank, models estimated HbA1c with an MAE of 0.34% and no estimation bias (MBE: 0.001%). In the prospective validation cohort, MAE for HbA1c was 1.96% at the baseline visit, and there was a substantial underestimation of HbA1c in model-produced estimates (MBE, baseline: -1.69%). Approximately 56% of model-produced estimations of HbA1c fell within 1 percentage point

of the baseline visit measurement, and 41.1% fell within 0.5 of a percentage point.

The model trained to identify patients with DM diagnoses had an AUROC of 0.73 in the UK Biobank and 0.76 in the prospective validation cohort. Using a threshold of a model-predicted probability of  $\geq 0.142$ , the model-produced classification of DM had a sensitivity of 0.67 and a specificity of 0.66 in the UK Biobank and a sensitivity of 0.80 and a specificity of 0.53 in the prospective validation cohort. In the UK Biobank, 7% of the cases classified as having DM by the model had a recorded diagnosis; in the prospective validation cohort, this value was 60%.

Using the HbA1c prediction model with thresholds of either 6.5% or 70% resulted in models with low sensitivity (0.25 and 0.11, respectively) but high specificity (0.94 and 1.00, respectively) for identifying patients with a DM diagnosis or suboptimally controlled diabetes.

#### Estimated glomerular filtration rate

Using CKD-EPI (2021), creatinine-based equations measured the mean eGFR in the UK Biobank training set at 89.3 ml/min/1.73 m<sup>2</sup> (SD 13.5) and 96.3 ml/min/1.73 m<sup>2</sup> (SD 16.8) in the Kenyan cohort at baseline. In the UK Biobank, there was a small difference between the MAE for model-produced estimates of creatinine-based eGFR (12.72 ml/min/1.73 m<sup>2</sup>) versus cystatin C-based eGFR (13.42 ml/min/1.73 m<sup>2</sup>). However, in the prospective validation cohort, eGFR estimates based on creatinine were markedly lower (baseline: 14.33 ml/min/1.73 m<sup>2</sup>) than those based on cystatin C (baseline: 22.36 ml/min/1.73 m<sup>2</sup>). While creatinine-based eGFR

models had minimal bias (MBE, baseline: 2.05 ml/min/1.73 m<sup>2</sup>), cystatin C-based eGFR models substantially overestimated lab-based values (baseline: 19.34 ml/min/1.73 m<sup>2</sup>). The model-produced estimates of eGFR fell within the 13 ml/min/1.73 m<sup>2</sup> tolerance limit for 58.6% of predictions based on creatinine and 35.3% of predictions based on cystatin C. Follow-up visit tolerance results were slightly lower than those observed during baseline.

#### *Additional exploratory results*

Machine learning models were additionally trained to estimate values for haematocrit, haemoglobin and red blood cells from UK Biobank data. Predictions in the validation cohort were comparable with those obtained on UK Biobank data; full results are presented in Table S1.

To investigate the model's explainability, occlusion and integrated gradient methods were applied. In brief, this post hoc analysis identified that anatomical features of importance for SBP prediction were predominantly blood vessels in hypertensive and normotensive participants. More detailed results of this analysis are presented in Figures S1 and S2.

#### *Safety*

This was a non-interventional study; two participants (0.7%) experienced a serious adverse event, and neither were deemed related to study procedures.

## 4 | DISCUSSION

Deep learning models trained on UK Biobank data to estimate clinical and laboratory parameters from retinal fundus images showed performance that was similar or marginally reduced in Kenyan study participants, despite substantial differences in race/ethnicity and a higher prevalence of DM and hypertension. The extreme values observed in the Kenyan cohort for some clinical parameters, particularly HbA1c, show unmet diagnosis and treatment needs in this population. With the routine collection of retinal images to screen for diabetic retinopathy, advances in this technology may hold potential to simultaneously ascertain related measures for patient care. This study also shows the feasibility of rapidly collecting high-quality laboratory, clinical and retinal image data needed to train and evaluate machine learning algorithms in resource-limited settings. All study procedures had a high ( $\geq 97\%$ ) level of completeness, and over 97% of participants had interpretable retinal images available for both eyes.

Additional studies showed the feasibility and challenges of data collection in similar settings. A 2-year study in Kenya's Nakuru County administered ophthalmological examinations to >4000 participants.<sup>12</sup> For 354 participants (10.2%), images were 'ungradable' for diabetic retinopathy.<sup>29</sup> Similar studies deploying and evaluating machine learning algorithms on retinal fundus images in Zambia and Thailand reported 0.3% and 14.5% ungradable images, respectively.<sup>9,30</sup> Taken alongside this study's results, there is evidence of the feasibility of a more extensive collection of similar data in future studies.

Lack of representation of certain groups in the data used to train a machine learning model can lead to poorer performance when the model is used in those groups, a phenomenon sometimes described as 'minority bias'.<sup>31</sup> When regression models are used to predict the risk of cardiovascular events, model parameters vary by race for key predictor variables,<sup>32</sup> meaning a model trained using data from primarily one racial group would lead to poorer risk estimation if deployed in populations comprised of other racial groups. For this reason, there have been calls for machine learning models to be adequately tested and trained on African populations, if that is where they will be used.<sup>9,30,33</sup> A strength of this study was the diversity between the initial training and subsequent validation datasets and the opportunity to directly compare results. Differences in the MBE metric showed continuous laboratory and vital sign measures were systematically over- or underestimated in the Kenyan prospective validation cohort, presenting opportunities for improving performance through model recalibration in the new population.

To ensure the prospective validation sample represented participants who could potentially benefit from machine learning-based diagnostic tools, enrolment criteria enriched the study population with participants with DM. This resulted in a study population that had a higher prevalence of obesity, hypertension and DM than the population in which the model was trained, in addition to substantial differences in race and ethnicity. Yet, for many endpoints, this study found a reasonably comparable model performance between the participants in Kenya and the UK Biobank held-out test, which is consistent with findings of other studies in which retinal image-based machine learning algorithms generalized well in new populations. An algorithm to detect diabetic retinopathy originally trained on a Singaporean population performed well when evaluated in Zambia.<sup>9</sup> Similarly, an algorithm trained on data from India and the United States performed well when prospectively deployed in a Thailand screening programme for diabetic retinopathy.<sup>30</sup> Both of these studies evaluated the performance of models for predicting referable versus non-referable diabetic retinopathy in type 2 DM, but did not explore predictions of continuous clinical biomarkers or other diagnoses. In contrast, another algorithm that predicted laboratory parameters from retinal images in retrospective cohorts had performance that varied across geography, particularly for creatinine estimation.<sup>15</sup>

Model performance was strong for parameters related to BP, including diagnosis of hypertension and prediction of SBP and DBP, with results approaching a level considered acceptable for a clinical diagnostic test for some parameters, i.e. AUROC >0.7.<sup>34</sup> In the post hoc analyses of model explainability, pixels near retinal blood vessels were identified as having importance to model predictions, consistent with previous findings<sup>14</sup> and providing biological plausibility to the results. Interestingly, prediction accuracy was not enhanced by taking a reading 15 min before or after an image compared with 60 min. As BP is labile and varies significantly throughout the day,<sup>35</sup> this finding implies predictions may rely to some extent on more subacute or chronic retinal features.

For predictions of continuous HbA1c levels, model performance was less accurate in the validation cohort, probably because of the

extreme values (up to 17.6%) observed in Kenyan patients, which were simply not present in the UK Biobank dataset. However, the prediction of the DM diagnosis was robust in the Kenyan cohort (AUROC 0.76) and patients with suboptimal glycaemic control (HbA1c >6.5%) could be detected with a high degree of specificity but low sensitivity. The model performance was also less accurate for eGFR estimation, and the variance between the test and the validation datasets was greater. This weaker performance is attributable to a less accurate estimation of creatinine and cystatin C, from which eGFR estimates were derived using the CKD-EPI equations.<sup>25</sup> Further work is required to distinguish between simple overfitting of the original models, and an inherent difference in the relationship between the retina and renal biomarkers in these populations.

This study has several limitations. First, the precision of the evaluation results from the prospective cohort is limited because of the small sample size. Second, in the prospective validation component of the study, most participants (281 of 301) were recruited from a single site, which may limit generalizability; however, the population was still substantially different than the one used to train the model. Third, by the time of publication, there had been substantial improvements in convolutional neural networks for image processing tasks beyond the Inception v3 architecture used in this paper. We expect further performance improvements will be possible as the deep learning methodology advances.

## 5 | CONCLUSION

This study represents a step towards leveraging machine learning to make early cardiovascular screening more accessible and sustainable. It shows proof-of-concept that retinal images, vital signs and laboratory measures can be reliably collected in resource-limited settings to support the training and evaluation of machine learning algorithms. The size of this dataset did not allow us to retrain the algorithm to account for the differences. Furthermore, there have been substantial improvements in convolutional neural networks for image processing tasks beyond the Inception v3 architecture used in this paper. We expect further performance improvements will be possible as the deep learning methodology advances.

As new models are developed for screening and diagnostic purposes, this study highlights that it will be essential to ensure they are calibrated to and have their performance validated in the populations where they will be used.

### AUTHOR CONTRIBUTIONS

VERP, AM, VS, TW, NS and FWS: conceptualization. GM, NO, VN and MSS: data curation. GM, TW, NS, NO and KR: formal analysis. VERP, AM and RE: funding acquisition. FWS, VS, VERP, VN, MSS and IA: investigation. GM, TW and NO: methodology. MZ, KK, TH-M, VN, MSS and IA: project administration. GM, TW and NS: software. RE, VN and MSS: supervision. VERP, VS, FWS, TW, NS, KR, KL and CM: writing— original draft. All authors: writing – review and editing. All authors had access to the data in this study and confirm accept responsibility to submit for publication.

### ACKNOWLEDGMENTS

This research has been conducted using the UK Biobank Resource. The authors gratefully acknowledge the contributions of Glen James, Andrew Cooper, Mishal Patel, Luke Markham, Matthew Pigg and Graham Morris. The authors would also like to thank AstraZeneca's Healthy Heart Africa programme for their support of this work (<https://www.astrazeneca.com/sustainability/access-to-healthcare/healthy-heart-africa.html>).

### FUNDING INFORMATION

AstraZeneca sponsored the study and company employees played active roles in study design, data analysis, study interpretation, writing of this manuscript, and in the decision to submit for publication.

### CONFLICT OF INTEREST STATEMENT

TW, VS, FWS, NS, GM, PF, AS, MM-S, KK, TMH, RE, AM and VERP are employees of AstraZeneca and hold AstraZeneca shares. KR, NO, KL and CM are employed by IQVIA; their involvement was funded by AstraZeneca. KR has previously worked for and holds equity in Google. FWS has previously worked for and is a shareholder in Novo Nordisk. CM is a stockholder in AZ, MindMed and J&J.

### PEER REVIEW

The peer review history for this article is available at <https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/dom.15587>.

### DATA AVAILABILITY STATEMENT

Data from the UK Biobank is available to investigators (please see <https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access>). Access to the code of the trained model is made available at <https://github.com/AstraZeneca>. Validation data is available through vivli.org in accordance with AstraZeneca's data sharing policy: <https://astrazenecagrouptrials.pharmacm.com/ST/Submission/Disclosure>.

### ORCID

Vikesh Selvarajah  <https://orcid.org/0009-0004-3662-7494>

Fredrik Wolfhagen-Sand  <https://orcid.org/0009-0007-7148-4144>

Victoria E. R. Parker  <https://orcid.org/0000-0003-2706-2669>

### REFERENCES

1. World Health Organization. The top 10 causes of death. 2020 Accessed February 21, 2022. <https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death>
2. World Health Organization. Global Health Estimates. 2019 Accessed February 21, 2022. <https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/gh-leading-causes-of-death>
3. World Health Organization. More than 700 million people with untreated hypertension. 2021.
4. Handler J. Clinical challenges in diagnosing and managing adult hypertension. *Cleve Clin J Med*. 2015;82(12 Suppl 2):S36-S41. doi:10.3949/ccjm.82.s2.06
5. World Health Organization. Global report on diabetes. WHO library cataloguing-in-publication data global report on diabetes. 2016.

6. Crisp N, Chen L. Global supply of health professionals. *N Engl J Med*. 2014;370(10):950-957. doi:10.1056/NEJMra1111610
7. Huang XM, Yang BF, Zheng WL, et al. Cost-effectiveness of artificial intelligence screening for diabetic retinopathy in rural China. *BMC Health Serv Res*. 2022;22(1):260. doi:10.1186/s12913-022-07655-6
8. Xie Y, Nguyen QD, Hamzah H, et al. Artificial intelligence for teleophthalmology-based diabetic retinopathy screening in a national programme: an economic analysis modelling study. *Lancet Digit Health*. 2020;2(5):e240-e249. doi:10.1016/S2589-7500(20)30060-1
9. Bellefleur V, Lim ZW, Lim G, et al. Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study. *Lancet Digit Health*. 2019;1(1):e35-e44. doi:10.1016/S2589-7500(19)30004-4
10. Gulshan 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(22):2402-2410. doi:10.1001/jama.2016.17216
11. Beede E, Baylor E, Hersch F, et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM; 2020:1-12. doi:10.1145/3313831.3376718
12. Bastawrous A, Mathenge W, Peto T, et al. The Nakuru eye disease cohort study: methodology & rationale. *BMC Ophthalmol*. 2014;14(1):60. doi:10.1186/1471-2415-14-60
13. 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(1):9432. doi:10.1038/s41598-020-65794-4
14. 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. doi:10.1038/s41551-018-0195-0
15. Rim TH, Lee G, Kim Y, et al. Prediction of systemic biomarkers from retinal photographs: development and validation of deep-learning algorithms. *Lancet Digit Health*. 2020;2(10):e526-e536. doi:10.1016/S2589-7500(20)30216-8
16. 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. doi:10.1371/journal.pone.0233166
17. Tham YC, Liu Y, Ting D, et al. Estimation of Haemoglobin A1c from retinal photographs via deep learning. *Invest Ophthalmol Vis Sci*. 2019;60(9):1456.
18. 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. doi:10.1109/ACCESS.2021.3052477
19. 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. doi:10.1038/s41551-021-00745-6
20. Dai G, He W, Xu L, et al. Exploring the effect of hypertension on retinal microvasculature using deep learning on east Asian population. *PLoS One*. 2020;15(3):e0230111. doi:10.1371/journal.pone.0230111
21. Lackland DT. Racial differences in hypertension: implications for high blood pressure management. *Am J Med Sci*. 2014;348(2):135-138. doi:10.1097/MAJ.0000000000000308
22. Lackland DT, Bachman DL, Carter TD, Barker DL, Timms S, Kohli H. The geographic variation in stroke incidence in two areas of the southeastern stroke belt. *Stroke*. 1998;29(10):2061-2068. doi:10.1161/01.STR.29.10.2061
23. Ford ES. Trends in mortality from all causes and cardiovascular disease among hypertensive and nonhypertensive adults in the United States. *Circulation*. 2011;123(16):1737-1744. doi:10.1161/CIRCULATIONAHA.110.005645
24. Barriada RG, Masip D. An overview of deep-learning-based methods for cardiovascular risk assessment with retinal images. *Diagnostics*. 2023;13(1):68. doi:10.3390/DIAGNOSTICS13010068
25. UK Biobank. 2023. Accessed February 21, 2023. <https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us>.
26. National Kidney Foundation. eGFR calculator. Accessed March 1, 2022. [https://www.kidney.org/professionals/kdoqi/gfr\\_calculator](https://www.kidney.org/professionals/kdoqi/gfr_calculator)
27. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. *2009 Conference on Computer Vision and Pattern Recognition*. IEEE; 2009:248-255. doi:10.1109/CVPR.2009.5206848
28. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. *2016 Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2016:2818-2826. doi:10.1109/CVPR.2016.308
29. Hansen MB, Abramoff MD, Folk JC, Mathenge W, Bastawrous A, Peto T. Results of automated retinal image analysis for detection of diabetic retinopathy from the Nakuru study, Kenya. *PLoS One*. 2015;10(10):e0139148. doi:10.1371/journal.pone.0139148
30. 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. doi:10.1016/S2589-7500(22)00017-6
31. Char DS, Shah NH, Magnus D. Implementing machine learning in health care – addressing ethical challenges. *N Engl J Med*. 2018;378(11):981-983. doi:10.1056/NEJMp1714229
32. Gijsberts CM, Groenewegen KA, Hoefler IE, et al. Race/ethnic differences in the associations of the Framingham risk factors with carotid IMT and cardiovascular events. *PLoS One*. 2015;10(7):e0132321. doi:10.1371/journal.pone.0132321
33. Mathenge WC. Artificial intelligence for diabetic retinopathy screening in Africa. *Lancet Digit Health*. 2019;1(1):e6-e7. doi:10.1016/S2589-7500(19)30009-3
34. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. *J Thorac Oncol*. 2010;5(9):1315-1316. doi:10.1097/JTO.0b013e3181ec173d
35. Kawano Y. Diurnal blood pressure variation and related behavioral factors. *Hypertens Res*. 2011;34(3):281-285. doi:10.1038/hr.2010.241

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** White T, Selvarajah V, Wolfhagen-Sand F, et al. Prediction of cardiovascular risk factors from retinal fundus photographs: Validation of a deep learning algorithm in a prospective non-interventional study in Kenya. *Diabetes Obes Metab*. 2024;1-10. doi:10.1111/dom.15587