Automated retinal image analysis systems to triage for grading of diabetic retinopathy: a large-scale, open-label, national screening programme in England

Alicja R Rudnicka, Royce Shakespeare, Ryan Chambers, Louis Bolter, John Anderson, Jiri Fajtl, Roshan A Welikala, Sarah A Barman, Abraham Olvera-Barrios, Laura Webster, Samantha Mann, Aaron Lee, Paolo Remagnino, Catherine Egan, Christopher G Owen, Adnan Tufail, on behalf of the ARIAS Research Group*

Lancet Digit Health 2025

Published Online
https://doi.org/10.1016/j.landig.2025.100914

See Comment https://doi.org/10.1016/j.landig.2025.100915

*Members are listed at the end of the article

School of Health and Medical Sciences, City St George's, University of London, London, UK (Prof A R Rudnicka PhD, R Shakespeare MSc, Prof C G Owen PhD); Diabetes and Endocrinology, Homerton Healthcare NHS Foundation Trust, London, UK (R Chambers BEng, L Bolter MSc, J Anderson MD); School of Computer Science and Mathematics, Kingston University, London, UK (J Fajtl PhD, R A Welikala PhD, Prof S A Barman PhD); NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, UK (A Olvera-Barrios MD, C Egan MD, Prof A Tufail MD); South East London Diabetic Eye Screening Programme, Guy's and St Thomas' NHS Foundation Trust, London, UK (L Webster, S Mann MD); Department of Ophthalmology, University of Washington, Seattle, WA, USA (A Lee MD); Department of Computer Science, Durham University, Durham, UK (Prof P Remagnino PhD)

Correspondence to:
Prof Alicja R Rudnicka, School of Health and Medical Sciences, City St George's, University of London, London SW17 0RE, UK
arudnick@citystgeorges.ac.uk

Summary

Background The global prevalence of diabetes is rising, alongside costs and workload associated with screening for diabetic eye disease (diabetic retinopathy). Automated retinal image analysis systems (ARIAS) could replace primary human grading of images for diabetic retinopathy. We evaluated multiple ARIAS in a real-life screening programme.

Methods Eight of 25 invited and potentially eligible CE-marked systems for diabetic retinopathy detection from retinal images agreed to participate. From 202 886 screening encounters at the North East London Diabetic Eye Screening Programme (between Jan 1, 2021, and Dec 31, 2022) we curated a database of 1·2 million images and sociodemographic and grading data. Images were manually graded by up to three graders according to a standard national protocol. ARIAS performance overall and by subgroups of age, sex, ethnicity, and index of multiple deprivation (IMD) were assessed against the reference standard, defined as the final human grade in the worst eye for referable diabetic retinopathy (primary outcome). Vendor algorithms did not have access to human grading data.

Findings Sensitivity across vendors ranged from 83·7% to 98·7% for referable diabetic retinopathy, from 96·7% to 99·8% for moderate-to-severe non-proliferative diabetic retinopathy, and from 95·8% to 99·5% for proliferative diabetic retinopathy. Sensitivity was largely consistent for moderate-to-severe non-proliferative and proliferative diabetic retinopathy by subgroups of age, sex, ethnicity, and IMD for all ARIAS. For mild-to-moderate non-proliferative diabetic retinopathy with referable maculopathy, sensitivity across vendors ranged from 79·5% to 98·3%, with greater variability across population subgroups. False positive rates for no observable diabetic retinopathy ranged from 4·3% to 61·4% and within vendors varied by 0·5 to 44 percentage points across population subgroups.

Interpretation ARIAS showed high sensitivity for medium-risk and high-risk diabetic retinopathy in a real-world screening service, with equitable performance across population subgroups. ARIAS could provide a cost-effective solution to deal with the rising burden of screening for diabetic retinopathy by safely triaging for human grading, substantially increasing grading capacity and rapid diabetic retinopathy detection.

Funding NHS Transformation Directorate, The Health Foundation, and The Wellcome Trust.

Introduction

Diabetes affects 1 in 10 adults globally, and rates worldwide have quadrupled over the past two decades, with approximately 537 million people currently having diabetes, projected to rise to 784 million people in 2045.[1] The majority of health-care costs for diabetes relate to complications.[2,3] Diabetic retinopathy is the main microvascular complication of diabetes and a leading cause of incident blindness among the working-age population.[4] The English NHS Diabetic Eye Screening Programme (DESP) carried out over 2·2 million diabetic eye screening appointments from April 1, 2017, to March 31, 2018.[5] The primary aim of this national programme is to "reduce the risk of sight loss amongst people with diabetes by the prompt identification and effective treatment if necessary of sight-threatening diabetic retinopathy ... with a long-term aim of preventing blindness in people with diabetes".[6,7] However, given the rising prevalence of diabetes, screening for diabetic retinopathy presents a huge challenge, requiring manual grading of over 12 million retinal images in England alone each year by up to three trained human graders (primary, secondary, and tertiary graders, reflecting level of experience and expertise) for the presence and severity of diabetic retinopathy. People with sight-threatening diabetic retinopathy are referred to hospital eye services for further assessment and potential treatment. Diabetic retinopathy screening is labour-intensive and costly, and requires ongoing training and quality assurance of human graders.[8] Several automated retinal image analysis systems (ARIAS), including artificial intelligence (AI) algorithms, can safely

Research in context

Evidence before this study

We searched PubMed from Jan 1, 2021, to Sept 26, 2025, to identify published evaluations and reviews of automated retinal image analysis systems (ARIAS) designed to fully or partly replace human grading for the detection of diabetic eye disease (diabetic retinopathy). We used the following search terms: ("Diabetic Retinopathy"[mh] OR "diabetic retinopathy"[tiab]) AND (("fundus photograph*"[tiab] OR "retinal image*"[tiab])) AND (("image processing, computer assisted/methods"[mh] OR "automated"[tiab] OR "AI"[tiab] OR "deep learning"[tiab] OR "machine learning"[tiab])) AND ("direct comparison"[tiab] OR "head-to-head"[tiab]). Most evaluations of ARIAS to date have largely relied on a single system applied to a specific dataset. The deployment of ARIAS in large-scale screening, such as national programmes, has been restricted globally. We identified two large-scale head-to-head evaluations of multiple vendors, but only one provided results with named vendors. Both studies were underpowered to assess equity and algorithmic fairness across different population subgroups.

Added value of this study

We present a large-scale, multi-vendor comparison of licensed ARIAS using a diverse dataset encompassing images with varying characteristics. The population in this study represents multiple ethnicities, a wide age range, varying levels of social deprivation, and the full spectrum of diabetic retinopathy severity, providing an accurate reflection of real-world health-care settings. To our knowledge, our evaluation is the most comprehensive conducted by a research team with no commercial interests in any particular algorithm. This study is unique in both its scale and the open reporting of the results for each algorithm. Moreover, our methodology could serve as a model for independent evaluation of algorithms in other real-world health-care settings.

Implications of all the available evidence

ARIAS showed similar sensitivity to human graders for detecting medium-risk and high-risk diabetic retinopathy in routine high-volume screening, with equitable precision in performance across diverse population subgroups. ARIAS have the potential to enhance grading capacity in diabetic eye disease screening by enabling rapid diabetic retinopathy detection and triage for human assessment. Multiple algorithms met the predefined standards for safety and workload reduction, strengthening the case for their deployment in large-scale diabetic eye screening. Our findings could stimulate further investment and innovation in combatting this globally prevalent blinding disease. This study provides a framework for transparent and reliable evaluation of clinical artificial intelligence systems in screening, providing valuable insights to guide health-care standards for ARIAS in a live screening setting.

and effectively triage patients into low-risk diabetic retinopathy (not requiring human grading) and medium-risk or high-risk diabetic retinopathy (requiring human grading),[9–11] which could considerably reduce the number of encounters requiring human grading.[8,10,12] ARIAS could substantially increase image-grading capacity and provide a cost-effective alternative to a purely human grading system.[8,10–12]

ARIAS that are trained and tested using retinal images from restricted demographic and ethnic groups might not be generalisable to other populations and could amplify differences between subgroups, thereby introducing or deepening bias. For example, ARIAS systems developed in predominantly White populations might not perform equally well in other ethnic groups,[13,14] given that differing levels of retinal pigmentation might affect image quality[12] and system performance, analogous to known errors in pulse oximetry with skin pigmentation.[15] As ARIAS enter clinical use, effective assessment of performance before, during, and after implementation in health care is increasingly important.

To avoid financial disinvestment in health service provision, mistrust in innovation, and public or stakeholder disengagement from technological advances, algorithmic evaluation must be transparent, trustworthy, and impartial. Furthermore, algorithmic fairness must be assessed across diverse population subgroups to ensure that systems meet predefined standards before they are deployed in the intended health-care settings.[13–15]

Most evaluations to date have used a single ARIAS on a specific dataset, making selection of the optimal ARIAS for deployment by health-care commissioners problematic. Head-to-head comparisons of ARIAS on the same large, real-life, diverse population data are needed to provide robust comparisons in the same computational environment. We outlined our methodology[16] for multivendor evaluation of equity and algorithmic fairness of clinical AI across different subgroups of the population, which complies with British Standards Institute BS 30440.[17] In this Article, we present a real-world evaluation done in a large, ethnically diverse screening programme in the North East London (NEL) DESP. We aimed to provide valuable, up-to-date information on ARIAS performance relevant to governmental, UK National Health Service (NHS), and lay stakeholders, and commercial vendors. Our intended use case for ARIAS is to safely triage medium-risk and high-risk diabetic retinopathy cases for human grading, thereby removing low-risk cases from the human grading queue. Our evaluation focused on assessment of ARIAS with (or pending) CE Class IIa medical device certification (ie, software licenced as a medical device).

Methods

Overview of the ARIAS evaluation process

Our methodology has been described in detail;[16] two invitations were sent to 25 (potential CE-marked) vendors (panel), giving them 15 months to submit algorithms for offline image processing to our secure server. Eight vendors accepted (appendix p 8).[16] ARIAS were developed using internationally recognised clinical scales for diabetic retinopathy grading.[18,19] All consecutive screening encounters with images between Jan 1, 2021, and Dec 31, 2022, from the NEL DESP were used to evaluate ARIAS performance. The NEL DESP, provided by Homerton Healthcare NHS Foundation Trust, offers annual (or biennial since October, 2023) screening to over 150 000 people aged 12 years or older living with diabetes. Patient consent for data collection and use is obtained upon entry to the NEL DESP and stored electronically. Uniquely, NEL DESP serves a population that comprises the three main ethnic groups in the UK (29% White, 17% Black, and 41% south Asian), showing the full spectrum of diabetic eye disease severity, with a wide age range. 4% of people in the NEL DESP have type 1 diabetes. The NEL DESP is located in one of the most deprived regions in the UK.[20] OptoMize software version 4.8 (NEC Software Solutions; Cheshire, UK) is used to manage and store screening and sociodemographic data.

Ethics

All data were managed according to UK NHS information governance requirements, in accordance with the principles of the Data Protection Act 2018. Methodologies adhere to the Homerton Healthcare NHS Foundation Trust Security Policy and Information Governance Policy, and General Data Protection Regulation principles relating to processing of personal data. The study has been approved by the Health Research Authority (IRAS project ID 265637). The NHS Health Research Authority toolkit identified that research ethics approval was not required for this project, as all data are pseudonymised and presented in aggregate form. A Data Protection Impact Assessment was submitted to Homerton Healthcare NHS Foundation Trust Information Governance Lead and approved by the information governance team in December, 2022.

Data sources

Retinal image capture followed UK National Screening Committee (NSC) protocols (two 45° field retinal images per eye centred on the optic disc and macula following mydriasis). A range of approved fundus cameras were used.[21]

The clinical care team uploaded pseudonymised images after removal of personal identifiable data and image metadata (circa 200 000 screening episodes with around 1·2 million images) to a trusted research environment for ARIAS processing that protects data privacy. A second dataset that included routinely recorded information (eg, age, sex, self-described ethnicity, visual acuity, index of multiple deprivation [IMD],[22] type of diabetes, duration of diabetes, and human grades following national grading classification)[23,24] was exported (using the same pseudonymised identifiers as for retinal images) to a separate location and not shared with ARIAS vendors.[16] Crucially, this real-life dataset of images from NEL DESP has not been used to develop or train any ARIAS.

Panel: Alphabetic list of manufacturers invited that potentially had CE mark (class IIa), with automated retinal image analysis system participation key

- Airdoc Inc: Retinal Image Intelligent Analysis Software
- Artelus Ltd: DRISTi 2.0*—participation key A
- Digital Diagnostics: IDX-DR
- Evolucare Technologies SAS: OphtAI 2.3*—participation key F
- EyeCheckup: Eyecheckup AI*—participation key C
- Eye2you AI: Eye2you
- Eyetelligence Pty Ltd: Eyetelligence Assure
- EyeNuk Inc: EyeArt v3.0.0*—participation key B
- EyRis: SELENA+
- Google Health/Verily Life Science LLC: Automated Retinal Disease Assessment
- iHealthScreen Inc: iPredict
- Intelligent Retinal Imaging Systems LLC: Intelligent Retinal Imaging System
- MONA.health: MONA*—participation key D
- NEC Software Solutions: NEC*—participation key E
- Optomed: Avenue AI
- Remido Innovative Solutions Pvt Ltd: Remidio*—participation key G
- RetinAI: RetinAI
- Retina-AI Health Inc: Retina-AI Galaxy
- RetinaLyza System A/S: RetinaLyze
- Retmarker SA: Retmarker*—participation key H
- Scottish Health Innovations Ltd: GradingM
- SigTuple Technologies: Drishti
- TeleMedC: DRLite
- Thirona: RetCAD
- VUNO: VUNO Med-Fundus AI

Participation key (A–H) refers to the identifier used for each system in the results tables and figures. *Fully participated in the evaluation.

See Online for appendix

Reference standards for evaluation of ARIAS

Retinal images are assessed by up to three trained human graders following UK NSC guidance standards for the presence and severity of diabetic retinopathy[23,24] and quality assurance.[25,26] The NEL DESP has 40 graders, working 2–24 h per week, and consistently delivers high-quality assurance. All primary grades with any diabetic retinopathy and 10% with no diabetic retinopathy receive higher-level grading.[23] The final human grade in the worst eye per encounter was used as the reference standard to assess ARIAS performance (appendix pp 5–7). Human graders in the NEL DESP have access to sociodemographic data and visual acuity, which if 6/12 (20/40) or worse, in the absence of amblyopia or macular degeneration, requires assessment of the anterior segment and macula.

The UK NSC grading classifications and commensurate Early Treatment Diabetic Retinopathy Study (ETDRS) retinopathy grade scores are[18] as follows: no observable diabetic retinopathy (R0), ETDRS scores 10, 14–15 inclusive; no observable diabetic maculopathy (M0); mild non-proliferative (background) diabetic retinopathy (R1),

	Overall (N=5773)		NEL DESP (N=4273)		SEL DESP (N=1500)	
	% (95% CI)	n/N	% (95% CI)	n/N	% (95% CI)	n/N
Ethnicity						
White	95·9 (94·9–96·7)	1930/2013	95·6 (94·4–96·6)	1325/1386	96·5 (94·7–97·8)	605/627
Black	97·1 (96·1–97·9)	1351/1391	97·3 (96·0–98·3)	835/858	96·8 (94·9–98·1)	516/533
South Asian	96·8 (95·8–97·6)	1473/1522	96·7 (95·6–97·5)	1394/1442	98·8 (93·2–100·0)	79/80
Other or missing	96·0 (94·4–97·2)	813/847	95·9 (94·0–97·4)	563/587	96·2 (93·0–98·1)	250/260
Age group, years						
>30	99·4 (97·7–99·9)	306/308	99·5 (97·4–100·0)	210/211	99·0 (94·4–100·0)	96/97
30 to <45	97·7 (96·6–98·6)	939/961	97·9 (96·5–98·8)	686/701	97·3 (94·5–98·9)	253/260
45 to <60	97·4 (96·7–98·0)	2287/2347	97·6 (96·7–98·3)	1646/1687	97·1 (95·5–98·3)	641/660
60 to <75	95·1 (94·0–96·1)	1580/1661	94·6 (93·2–95·7)	1215/1285	97·1 (94·8–98·5)	365/376
≥75	91·7 (89·0–94·0)	455/496	92·5 (89·5–95·0)	360/389	88·8 (81·2–94·1)	95/107
Overall	96·4 (95·9–96·9)	5567/5773	96·3 (95·7–96·9)	4117/4273	96·7 (95·6–97·5)	1450/1500

n=number of R3A encounters primary graders classified as referable diabetic retinopathy. N=total number of R3A encounters per subgroup. NEL DESP=North East London Diabetic Eye Screening Programme. SEL DESP=South East London Diabetic Eye Screening Programme.

Table 1: **Data from NEL and SEL DESPs for encounters with final human outcome grade of active proliferative diabetic retinopathy (R3A) in worst eye between March 1, 2014, and Dec 31, 2022**

ETDRS scores 20–35 inclusive; ungradable images (U); moderate-to-severe non-proliferative (pre-proliferative) diabetic retinopathy (R2), ETDRS scores 43–53 inclusive; any diabetic maculopathy (M1); and proliferative diabetic retinopathy (R3), ETDRS scores ≥61.[23,24] In the English NHS DESP, retinopathy grades R0M0 and R1M0 are non-referable diabetic retinopathy and grades M1, R2, and R3 are referable diabetic retinopathy. Patients with referable diabetic retinopathy or ungradable images are reviewed by the referral outcome grader to confirm referral to a hospital eye service for specialist assessment and treatment if required. Patients with non-referable retinopathy receive an invitation for rescreening per protocol.[23]

Standards for detection of referable diabetic retinopathy

English NHS DESP quality assurance standards for human graders[25] examine human graders' ability to correctly refer patients compared with the reference standard. Human grader sensitivity greater than 85% for referable diabetic retinopathy is defined as adequate; 85% or lower is inadequate. Currently, there is no additional standard for the most severe retinopathy (proliferative diabetic retinopathy [R3]). To minimise unnecessary referrals to the hospital eye service, specificity is set to more than 80% (ie, to correctly not refer in the absence of referable diabetic retinopathy). To contextualise ARIAS performance against the reference standard, we compared ARIAS with primary graders. This comparison will be biased in favour of primary graders because of the aforementioned selective higher-level grading described.

Statistical analysis

We outlined our sample size calculation for equity in performance precision,[16] as defined by the lower bound of the 95% CIs for detection of the most serious and rarest outcome, proliferative diabetic retinopathy (R3). We anticipated ARIAS sensitivities of 90% or greater for R3. A priori, we agreed that a 1·0 percentage point difference or less in the lower bound of the 95% CI was an equitable precision standard for specified detection rates for R3 across the three main ethnic groups, which required curation of 200 000 consecutive encounters.[16]

ARIAS outcomes were merged via the pseudonymised identifier, with the DESP dataset containing the human grades and sociodemographic data. A positive ARIAS test is indicative of some form of diabetic retinopathy or not assessable image or technical failure by the algorithm (requiring human review to ascertain the reason for technical failure, eg, cataract obscuring retina actions slit-lamp biomicroscopy to visualise the retina) and ARIAS negative test is indicative of disease absence. ARIAS sensitivity (detection rate), false positive rates (ie, 100% minus-specificity as %), positive and negative predictive values, and likelihood ratios were calculated against the human reference standard of presence of referable diabetic retinopathy (yes *vs* no) and for each reference standard diabetic retinopathy grade by ethnicity (White, Black, south Asian, or other or unknown), age (<30 years, 30 years to <45 years, 45 years to <60 years, 60 years to <75 years, and ≥75 years), sex, diabetes type, and IMD quintiles. Age groups, rather than continuous age, were used to compare performance in those younger than 30 years versus older adults, in whom eye diseases like cataracts might affect retinal visibility and impact both human and ARIAS performance. Throughout, 95% CIs are logit-transformed or binomial exact for values of 100%. Given the large sample size, algorithmic fairness was assessed by graphically examining and summarising the absolute percentage differences in ARIAS and primary human grader performance across population subgroups compared with the reference standard.

ARIAS screen-positive rates estimated the proportion of all screening encounters that would require human grading if ARIAS were to be implemented as a first-pass triage before human grading. Sensitivity analyses examined

	White (N=64 023)	Black (N=34 058)	South Asian (N=78 415)	Other or unknown ethnicity (N=24 942)	Age <30 years (N=4219)	Age 30 to <45 years (N=21 871)	Age 45 to <60 years (N=66 941)	Age 60 to <75 years (N=74 231)	Age ≥75 years (N=34 176)	Overall (N=201 438)
A—Artelus Ltd: DRISTi 2.0										
No referable diabetic retinopathy	54·4 (54·0–54·8)	37·8 (37·2–38·3)	38·5 (38·2–38·9)	41·8 (41·2–42·5)	27·2 (25·8–28·7)	25·4 (24·8–26·0)	34·0 (33·6–34·4)	49·4 (49·0–49·8)	66·4 (65·9–67·0)	44·0 (43·8–44·2)
Referable diabetic retinopathy (including ungradable)	85·5 (84·7–86·3)	79·0 (78·0–80·0)	83·2 (82·5–83·8)	84·0 (82·7–85·2)	77·6 (73·8–81·1)	84·3 (83·0–85·7)	84·3 (83·5–85·0)	82·7 (82·0–83·4)	81·4 (80·3–82·4)	83·0 (82·6–83·5)
Referable diabetic retinopathy (excluding ungradable)	91·0 (90·3–91·7)	83·9 (82·8–84·9)	86·4 (85·7–87·0)	88·0 (86·8–89·1)	79·4 (75·5–82·9)	85·9 (84·6–87·2)	86·8 (86·0–87·4)	87·8 (87·1–88·5)	88·9 (87·8–90·0)	87·2 (86·8–87·6)
B—EyeNuk Inc: EyeArt v3.0.0										
No referable diabetic retinopathy	35·8 (35·4–36·2)	48·1 (47·5–48·6)	39·7 (39·4–40·1)	40·0 (39·3–40·6)	44·6 (43·0–46·2)	32·4 (31·7–33·1)	32·9 (32·6–33·3)	40·6 (40·2–41·0)	56·2 (55·6–56·8)	39·8 (39·6–40·1)
Referable diabetic retinopathy (including ungradable)	96·7 (96·2–97·0)	98·4 (98·0–98·7)	98·1 (97·9–98·4)	98·2 (97·7–98·6)	98·9 (97·5–99·6)	98·4 (97·9–98·8)	98·4 (98·1–98·6)	97·6 (97·3–97·9)	96·8 (96·3–97·2)	97·8 (97·6–98·0)
Referable diabetic retinopathy (excluding ungradable)	97·8 (97·4–98·1)	98·5 (98·1–98·8)	98·3 (98·0–98·5)	98·6 (98·1–99·0)	99·2 (97·9–99·8)	98·6 (98·0–99·0)	98·6 (98·3–98·8)	98·0 (97·7–98·3)	97·5 (96·9–98·0)	98·2 (98·1–98·4)
C—EyeCheckup: Eyecheckup AI										
No referable diabetic retinopathy	62·4 (62·0–62·8)	56·7 (56·1–57·2)	60·3 (59·9–60·7)	58·2 (57·6–58·9)	47·6 (46·0–49·3)	53·2 (52·5–53·9)	57·9 (57·5–58·3)	61·4 (61·0–61·8)	68·1 (67·6–68·6)	60·1 (59·9–60·4)
Referable diabetic retinopathy (including ungradable)	91·7 (91·1–92·3)	91·1 (90·4–91·8)	93·3 (92·9–93·7)	91·8 (90·8–92·7)	96·2 (94·1–97·6)	95·8 (95·0–96·5)	94·7 (94·3–95·1)	91·0 (90·5–91·5)	88·3 (87·4–89·1)	92·3 (92·0–92·6)
Referable diabetic retinopathy (excluding ungradable)	96·4 (95·9–96·9)	94·2 (93·5–94·9)	96·1 (95·7–96·5)	95·8 (95·0–96·5)	97·3 (95·4–98·6)	97·1 (96·4–97·7)	96·7 (96·4–97·1)	95·0 (94·6–95·5)	93·7 (92·8–94·5)	95·8 (95·5–96·0)
D—MONA.health: MONA										
No referable diabetic retinopathy	10·0 (9·7–10·2)	16·3 (15·9–16·7)	15·1 (14·8–15·4)	14·5 (14·0–15·0)	14·4 (13·3–15·6)	11·4 (11·0–11·9)	12·3 (12·1–12·6)	13·7 (13·5–14·0)	16·9 (16·5–17·3)	13·5 (13·4–13·7)
Referable diabetic retinopathy (including ungradable)	69·3 (68·3–70·3)	76·2 (75·1–77·2)	77·9 (77·1–78·6)	77·8 (76·4–79·1)	87·4 (84·2–90·1)	86·4 (85·1–87·6)	82·7 (82·0–83·5)	73·2 (72·4–74·0)	59·1 (57·8–60·4)	75·2 (74·8–75·7)
Referable diabetic retinopathy (excluding ungradable)	82·5 (81·6–83·5)	84·0 (82·9–85·0)	84·6 (83·9–85·2)	86·6 (85·3–87·8)	90·2 (87·2–92·7)	88·5 (87·2–89·6)	86·7 (86·0–87·4)	82·9 (82·1–83·7)	76·2 (74·7–77·7)	84·2 (83·7–84·7)
E—NEC Software Solutions: NEC										
No referable diabetic retinopathy	70·5 (70·1–70·9)	69·3 (68·8–69·9)	68·9 (68·6–69·3)	68·9 (68·3–69·5)	83·5 (82·3–84·7)	66·3 (65·7–67·0)	65·8 (65·4–66·2)	69·4 (69·0–69·7)	77·5 (77·0–78·0)	69·5 (69·3–69·7)
Referable diabetic retinopathy (including ungradable)	95·4 (94·9–95·8)	95·0 (94·4–95·5)	96·6 (96·3–96·9)	96·1 (95·5–96·7)	98·7 (97·3–99·5)	98·5 (97·9–98·9)	97·7 (97·4–98·0)	95·0 (94·6–95·4)	92·8 (92·1–93·5)	95·9 (95·7–96·1)

(Table 2 continues on next page)

	White (N=64 023)	Black (N=34 058)	South Asian (N=78 415)	Other or unknown ethnicity (N=24 942)	Age <30 years (N=4219)	Age 30 to <45 years (N=21 871)	Age 45 to <60 years (N=66 941)	Age 60 to <75 years (N=74 231)	Age ≥75 years (N=34 176)	Overall (N=201 438)
(Continued from previous page)										
Referable diabetic retinopathy (excluding ungradable)	99·1 (98·9–99·3)	98·2 (97·8–98·5)	98·6 (98·3–98·8)	98·9 (98·4–99·2)	99·8 (98·8–100·0)	99·0 (98·5–99·3)	98·9 (98·7–99·1)	98·6 (98·3–98·8)	97·9 (97·3–98·3)	98·7 (98·5–98·8)
F—Evolucare Technologies SAS: OphtAI 2.3										
No referable diabetic retinopathy	25·2 (24·8–25·5)	26·7 (26·2–27·2)	25·5 (25·2–25·9)	24·6 (24·0–25·2)	24·2 (22·9–25·7)	20·1 (19·6–20·7)	21·5 (21·2–21·9)	26·1 (25·7–26·4)	35·8 (35·2–36·3)	25·5 (25·3–25·7)
Referable diabetic retinopathy (including ungradable)	82·6 (81·8–83·5)	82·3 (81·3–83·3)	84·4 (83·8–85·1)	83·9 (82·6–85·0)	95·0 (92·8–96·7)	91·0 (89·9–92·0)	88·4 (87·7–89·0)	81·2 (80·5–82·0)	74·3 (73·1–75·4)	83·5 (83·0–83·9)
Referable diabetic retinopathy (excluding ungradable)	92·4 (91·7–93·0)	87·4 (86·4–88·3)	89·3 (88·7–89·8)	90·2 (89·1–91·3)	97·3 (95·4–98·6)	92·6 (91·6–93·6)	91·4 (90·8–92·0)	88·5 (87·8–89·1)	85·2 (83·9–86·4)	89·8 (89·4–90·1)
G—Remido Innovative Solutions Pvt Ltd: Remidio										
No referable diabetic retinopathy	17·2 (16·8–17·5)	21·8 (21·3–22·3)	19·7 (19·4–20·0)	19·1 (18·6–19·6)	20·4 (19·2–21·8)	13·5 (13·0–14·0)	15·5 (15·2–15·8)	19·9 (19·6–20·2)	28·3 (27·8–28·9)	19·1 (19·0–19·3)
Referable diabetic retinopathy (including ungradable)	77·8 (76·8–78·7)	79·6 (78·5–80·6)	81·9 (81·2–82·6)	81·3 (80·0–82·5)	91·2 (88·4–93·5)	88·0 (86·7–89·2)	85·1 (84·4–85·8)	78·6 (77·8–79·3)	70·0 (68·8–71·2)	80·3 (79·8–80·7)
Referable diabetic retinopathy (excluding ungradable)	87·8 (86·9–88·6)	85·3 (84·3–86·3)	87·1 (86·5–87·7)	87·9 (86·7–89·1)	92·3 (89·5–94·5)	89·7 (88·5–90·8)	88·4 (87·7–89·0)	86·3 (85·5–87·0)	82·2 (80·8–83·5)	87·0 (86·6–87·4)
H—Retmarker SA: Retmarker										
No referable diabetic retinopathy (R0, R1)	32·5 (32·1–32·9)	23·6 (23·1–24·1)	19·4 (19·1–19·7)	21·9 (21·3–22·4)	35·4 (33·8–36·9)	17·3 (16·8–17·9)	17·9 (17·6–18·2)	25·2 (24·8–25·5)	40·4 (39·8–41·0)	24·6 (24·4–24·9)
Referable diabetic retinopathy (including ungradable)	81·4 (80·5–82·2)	74·2 (73·1–75·3)	78·4 (77·7–79·1)	79·7 (78·4–81·0)	86·6 (83·4–89·4)	84·2 (82·8–85·6)	82·2 (81·4–83·0)	76·5 (75·7–77·3)	72·3 (71·1–73·4)	78·5 (78·1–79·0)
Referable diabetic retinopathy (excluding ungradable)	87·6 (86·7–88·4)	80·3 (79·1–81·4)	82·8 (82·1–83·5)	84·7 (83·3–85·9)	89·8 (86·7–92·4)	85·7 (84·4–87·0)	84·6 (83·9–85·4)	82·4 (81·6–83·2)	81·9 (80·5–83·2)	83·7 (83·2–84·1)

N=201 438 encounters from 126 365 people living with diabetes. Data show the proportion of encounters classified as test-positive by ARIAS (95% CIs). The primary outcome is defined by the reference standard, which is the final human grade in the worst eye. ARIAS test positive is indicative of some level of diabetic retinopathy or not assessable or technical failure by the algorithm. The overall column percentages represent overall ARIAS test positives for the primary outcome. Rows labelled referable diabetic retinopathy represent ARIAS sensitivity (detection rates) for referable diabetic retinopathy (moderate-to-severe non-proliferative [pre-proliferative] retinopathy [R2], diabetic maculopathy [M1], proliferative retinopathy [R3]). Rows labelled no referable diabetic retinopathy represent ARIAS false positives for no referable diabetic retinopathy. ARIAS=Automated Retinal Image Analysis System. R0=No observable diabetic retinopathy. R1=mild non-proliferative diabetic retinopathy.

***Table 2:* ARIAS performance for vendors A–H for the primary outcome (referable diabetic retinopathy (yes *vs* no); overall and by ethnicity and age group)**

ARIAS performance if encounters with visual acuity of 6/12 or worse were also classified as screen positive. Findings for ARIAS that had another CE-approved operating threshold are shown in the appendix (pp 36–47).

Stata version 18 was used for statistical analyses.

External validation of NEL DESP primary graders for active proliferative diabetic retinopathy

We examined primary grader performance for the most serious retinopathy, active proliferative diabetic retinopathy (R3A), in two sociodemographically diverse screening centres, NEL DESP and South East London (SEL) DESP.[10] Both centres screen over 100 000 people each year. Among those with a final human grade in the worst eye of R3A over an 8-year period (March 1, 2014, to Dec 31, 2022), the proportion that primary graders correctly defined as referable diabetic retinopathy in each centre by subgroups of age, sex, and ethnicity was determined. P values for inclusion of interaction terms between subgroups and centre were tested using logistic regression likelihood ratio tests.

Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

We recorded the proportion of encounters with final human grade in the worst eye of R3A (over an 8-year period in two centres) that primary graders classified as referable (table 1). Primary graders correctly graded 96·4% (95% CI 95·9–96·9) of R3A as referable diabetic retinopathy. The proportion of R3A encounters classified as referable by primary graders declined with increasing patient age. We found no evidence to suggest that primary grader classifications by subgroups (age, sex, and ethnicity) differed across centres (p interactions >0·20 in all instances).

In total, 202 886 encounters (126 365 people, 1·2 million images, median six images per screening encounter) were processed through eight ARIAS algorithms; 201 438 encounters had complete grading data for analysis (appendix p 10). Screened participants' mean age was 60·5 years (SD 14·3). 95 070 (47%) of 201 438 screened participants were female and 106 368 (53%) were male. 181 172 (93%) of 201 438 screened participants had type 2 diabetes. 64 023 (32%) of 201 438 screened participants were White, 34 058 (17%) were Black, 78 415 (39%) were south Asian, and 24 942 (12%) were of other or unknown ethnicity (1915 [0·5%] missing data). 30–48% of each ethnic group were from the lowest quintile of material circumstance, as defined by IMD, and 116 531 (58%) of 201 438 screened participants had diabetes diagnosed less than 10 years ago (appendix p 9). We had good representation of all grades of diabetic retinopathy, with 24 479 referable cases (tables 2, 3; appendix p 11).

For referable diabetic retinopathy (grades M1, R2, and R3 combined), ARIAS sensitivities for the eight vendors (A–H)

	R0M0 (N=131 669)	R1M0 (N=39 668)	R1M1 (N=18 620)	R2M0/M1 (N=3881)	R2M0 (N=999)	R2M1 (N=2882)	R3M0/M1 (N=1978)	R3M0 (N=826)	R3M1 (N=1152)	Referable diabetic retinopathy (N=24 479)
A—Artelus Ltd: DRISTi 2.0	38·9 (38·7–39·2)	60·8 (60·3–61·3)	83·7 (83·2–84·3)	98·0 (97·5–98·4)	93·8 (92·1–95·2)	99·4 (99·1–99·7)	98·7 (98·1–99·1)	97·2 (95·9–98·2)	99·7 (99·2–99·9)	87·2 (86·8–87·6)
B—EyeNuk Inc: EyeArt v3.0.0	26·4 (26·1–26·6)	84·5 (84·1–84·8)	97·8 (97·6–98·0)	99·8 (99·6–99·9)	99·3 (98·6–99·7)	100·0 (99·8–100·0)	99·4 (98·9–99·7)	98·5 (97·5–99·2)	100·0 (99·7–100·0)	98·2 (98·1–98·4)
C—EyeCheckup: Eyecheckup AI	50·6 (50·3–50·8)	91·9 (91·6–92·2)	94·8 (94·5–95·1)	99·6 (99·4–99·8)	99·1 (98·3–99·6)	99·8 (99·6–99·9)	97·5 (96·7–98·1)	94·9 (93·2–96·3)	99·3 (98·6–99·7)	95·8 (95·5–96·0)
D—MONA.health: MONA	4·3 (4·2–4·4)	44·1 (43·6–44·6)	79·9 (79·3–80·5)	98·4 (98·0–98·8)	95·1 (93·6–96·3)	99·6 (99·3–99·8)	96·8 (95·9–97·5)	93·1 (91·2–94·7)	99·5 (98·9–99·8)	84·2 (83·7–84·7)
E—NEC Software Solutions: NEC	61·4 (61·1–61·6)	96·6 (96·4–96·7)	98·3 (98·1–98·5)	99·8 (99·7–99·9)	99·5 (98·8–99·8)	100·0 (99·8–100·0)	99·5 (99·1–99·8)	98·9 (97·9–99·5)	100·0 (99·7–100·0)	98·7 (98·5–98·8)
F—Evolucare Technologies SAS: OphtAI 2.3	13·2 (13·0–13·4)	66·3 (65·9–66·8)	87·3 (86·8–87·8)	98·5 (98·1–98·9)	96·1 (94·7–97·2)	99·3 (99·0–99·6)	95·8 (94·8–96·6)	92·5 (90·5–94·2)	98·2 (97·2–98·9)	89·8 (89·4–90·1)
G—Remido Innovative Solutions Pvt Ltd: Remidio	8·6 (8·4–8·7)	54·3 (53·8–54·8)	83·5 (82·9–84·0)	98·9 (98·5–99·2)	96·9 (95·6–97·9)	99·6 (99·3–99·8)	97·1 (96·2–97·8)	93·8 (92·0–95·4)	99·4 (98·8–99·8)	87·0 (86·6–87·4)
H—Retmarker SA: Retmarker	18·2 (18·0–18·4)	46·1 (45·6–46·6)	79·5 (78·9–80·1)	96·7 (96·1–97·2)	89·9 (87·9–91·7)	99·1 (98·6–99·4)	97·7 (96·9–98·3)	95·0 (93·3–96·4)	99·6 (99·0–99·9)	83·7 (83·2–84·1)

N=201 438 encounters from 126 365 people living with diabetes. Data show the proportion of encounters classified as test-positive by ARIAS (95% CIs). Secondary outcomes are defined by the reference standard, which is the final human grade in the worst eye. ARIAS=Automated Retinal Image Analysis System. M0=no observable diabetic maculopathy. M1=diabetic maculopathy. R0=no observable diabetic retinopathy. R1=mild non-proliferative retinopathy. R2=moderate-to-severe non-proliferative (pre-proliferative) retinopathy. R3=proliferative retinopathy.

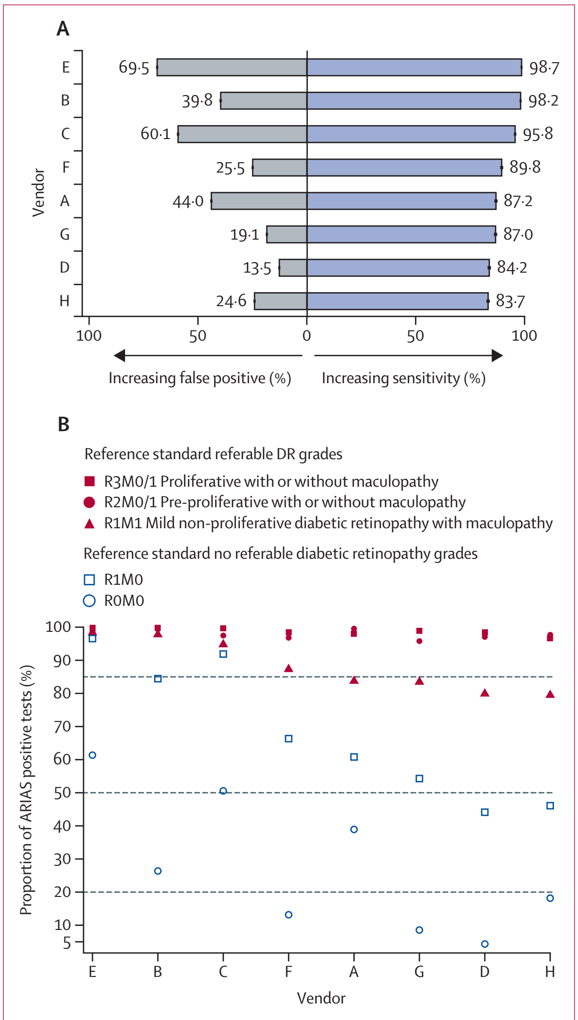
Table 3: **ARIAS performance for vendors A–H for the secondary outcomes (diabetic retinopathy grade; overall only)**

ranged from 83·7% to 98·7%. 13·5% to 69·5% of cases without referable diabetic retinopathy (grades R1M0 and R0M0) were also classified as positive by ARIAS (figure 1A; appendix p 12). Algorithms with a short bar to the left of 0 and a long bar to the right show the most discrimination for the primary outcome (figure 1A). Sensitivity for referable diabetic retinopathy was largely governed by the greater frequency of R1M1 cases (tables 2, 3). Additionally, variation by subgroups (tables 2, 3; appendix pp 13–14) was driven by variability in ARIAS sensitivity for R1M1 not R2, R3, or R3A. As expected, inclusion of encounters with final human grade of ungradable resulted in lower detection rates for all vendors.

ARIAS positive test rates for no referable diabetic retinopathy varied by population subgroups for all vendors (table 2). However, the absolute percentage differences were typically less than 10 percentage points, with larger variation between the youngest and oldest age groups (appendix pp 15–16).

ARIAS false positives across vendors ranged from 4·3% to 61·4% for R0M0 (no observable diabetic retinopathy with no observable diabetic maculopathy; ie, specificity 95·7% to 38·6%). ARIAS sensitivity for mild non-proliferative diabetic retinopathy without diabetic maculopathy (R1M0) ranged from 44·1% to 96·6% (table 3; figure 1B). Sensitivities for mild non-proliferative diabetic retinopathy with diabetic maculopathy (R1M1) ranged from 79·5% to 98·3% across vendors, and within ARIAS sensitivity varied across population subgroups by typically less than 5% (up to 16% variation across age groups; appendix pp 13, 20–27). For moderate-to-severe non-proliferative diabetic retinopathy (R2; ETDRS 43–53) or proliferative diabetic retinopathy (R3; ETDRS ≥61), with or without diabetic maculopathy, ARIAS sensitivities ranged from 96·7% (95% CI 96·1–97·2) to 99·8% (99·7–99·9) and from 95·8% (94·8–96·6) to 99·5% (99·1–99·8), respectively (table 3; appendix pp 20–27). For R2 and R3, within ARIAS variation in sensitivity across subgroups was typically less than 2 percentage points, with slightly higher variation (up to 7%) associated with the older age groups (figures 2, 3; appendix pp 20–27). Primary graders' sensitivity declined with increasing patient age for grades R2, R3, and R3A, and for some ARIAS (figure 3).

Overall screen positive rates ranged from 23% to 74% (figure 1C). Positive predictive values ranged from 17% to 47%, indicating that, among tests, between approximately 1 in 5 to 1 in 2 would have referable diabetic retinopathy as per the reference standard. For all vendors, negative predictive values were above 96%. Positive likelihood ratios for the ARIAS test ranged from 1·42 to 6·22 and negative likelihood ratios ranged from 0·03 to 0·23.

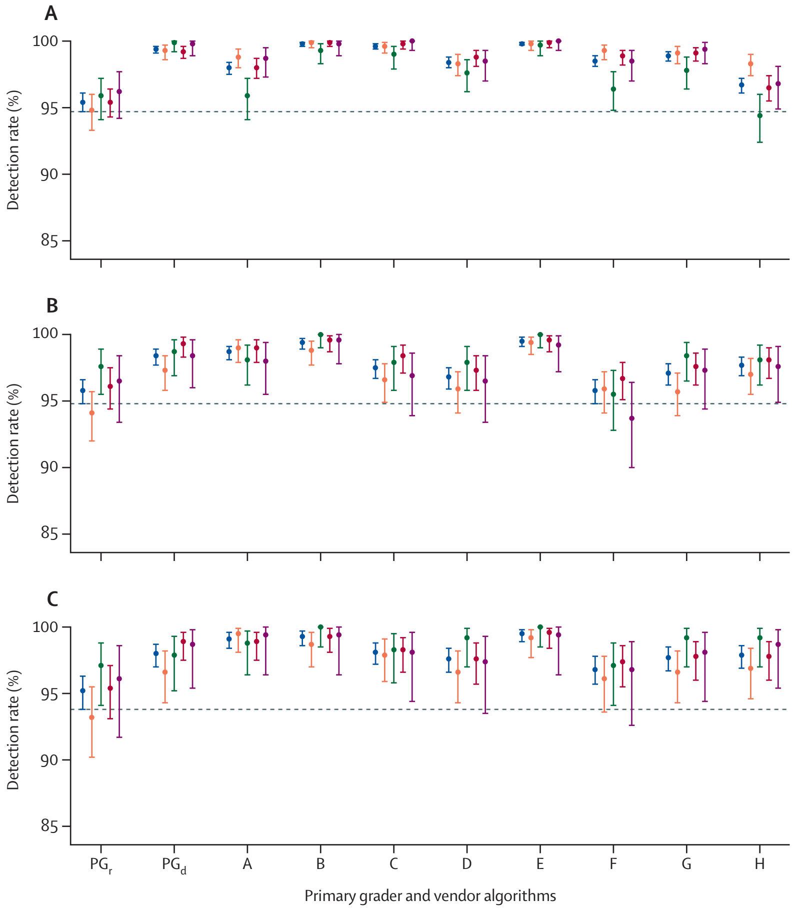


C

Vendor	% screen positive (n)	% positive predictive value (95% CI)	% negative predictive value (95% CI)
A	49·8 (100 391)	22·1 (21·8–22·3)	96·8 (96·7–96·9)
B	48·5 (97 671)	26·1 (25·8–26·3)	99·6 (99·5–99·6)
C	64·9 (130 812)	18·5 (18·3–18·7)	98·5 (98·4–98·6)
D	22·8 (45 860)	47·0 (46·6–47·5)	97·5 (97·4–97·5)
E	73·5 (147 957)	16·9 (16·7–17·1)	99·4 (99·3–99·4)
F	34·1 (68 779)	33·5 (33·1–33·8)	98·1 (98·0–98·1)
G	28·3 (56 956)	39·4 (39·0–39·8)	97·8 (97·7–97·8)
H	32·7 (65 862)	32·7 (32·3–33·0)	97·0 (96·9–97·1)

***Figure 1:* Performance metrics for ARIAS vendors A–H for the primary and secondary outcomes, as defined by the reference standard of final human grade in the worst eye**

Positive ARIAS test is indicative of some level or form of DR or not assessable or technical failure by the algorithm. (A) ARIAS test positives for the primary outcome of referable DR (reference standard DR grades R2, M1, or R3) are shown as sensitivity (detection rates) and ARIAS test positives for the primary outcome of no referable DR (reference standard DR grades R0M0 and R1M0) are shown as false positives. (B) Proportion of ARIAS positive tests for the secondary outcomes (ie, each reference standard DR grade separately). For DR grades R1, M1, R2, and R3, percentage values represent the sensitivity for each grade. For reference standard R0M0, percentage values represent the false positive rate for those without observable DR or maculopathy. (C) Overall screen positive rate for ARIAS A to H and the positive predictive value and negative predictive value for each ARIAS for referable DR as per the reference standard. ARIAS=automated retinal image analysis systems. DR=diabetic retinopathy. M0=no observable diabetic maculopathy. M1=diabetic maculopathy. n=number of encounters. R0=no observable diabetic retinopathy. R1=mild non-proliferative retinopathy. R2=moderate-to-severe non-proliferative (pre-proliferative) retinopathy. R3=proliferative retinopathy.



Ethnic group	Primary grader classification (95% CI)	
	% referable diabetic retinopathy	% diabetic retinopathy present
	R2M0 + R2M1	
Overall	**95·4 (94·7–96·1)**	**99·4 (99·1–99·6)**
White	94·8 (93·3–96·0)	99·3 (98·6–99·7)
Black	95·9 (94·1–97·2)	99·9 (99·2–100)
South Asian	95·4 (94·3–96·4)	99·2 (98·7–99·6)
Other/missing	96·2 (94·2–97·7)	99·8 (98·9–100)

	R3M0 + R3M1	
Overall	**95·8 (94·8–96·6)**	**98·4 (97·7–98·9)**
White	94·1 (92·0–95·7)	97·3 (95·8–98·4)
Black	97·6 (95·5–98·9)	98·7 (96·9–99·6)
South Asian	96·1 (94·4–97·5)	99·3 (98·3–99·8)
Other/missing	96·5 (93·4–98·4)	98·4 (96·0–99·6)

	R3AM0 + R3AM1	
Overall	**95·2 (93·8–96·3)**	**98·0 (97·0–98·7)**
White	93·2 (90·2–95·5)	96·6 (94·3–98·2)
Black	97·1 (94·1–98·8)	97·9 (95·2–99·3)
South Asian	95·4 (93·1–97·1)	98·9 (97·5–99·6)
Other/missing	96·1 (91·7–98·6)	98·7 (95·4–99·8)

Figure 2: **Performance for PGs and ARIAS (vendor algorithms A–H) by ethnic group**

Numerical values represent the proportion of encounters with specified reference standard diabetic retinopathy grades that were classified as referable diabetic retinopathy (PG$_r$) or diabetic retinopathy present (PG$_d$) by PG. For vendor algorithms, A to H values represent the proportion of encounters that were ARIAS test positive. Error bars are 95% CIs. (A) Reference standard diabetic retinopathy grade R2; moderate-to-severe non-proliferative diabetic retinopathy. (B) Reference standard diabetic retinopathy grade R3; proliferative diabetic retinopathy including those with or without diabetic maculopathy (grades M0 and M1). (C) Reference standard diabetic retinopathy grade R3A; active proliferative diabetic retinopathy, including those with or without diabetic maculopathy (grades M0 and M1). Blue indicates all ethnicities combined, orange indicates White ethnicity, green indicates Black ethnicity; red indicates south Asian ethnicity; and purple indicates all other ethnic groups. The horizontal dashed line is the lower 95% CI limit for PG$_r$ for all subgroups combined. ARIAS=automated retinal image analysis systems. M0=no observable diabetic maculopathy. M1=diabetic maculopathy. PG=primary grader. R2=moderate-to-severe non-proliferative (pre-proliferative) retinopathy. R3=proliferative retinopathy.

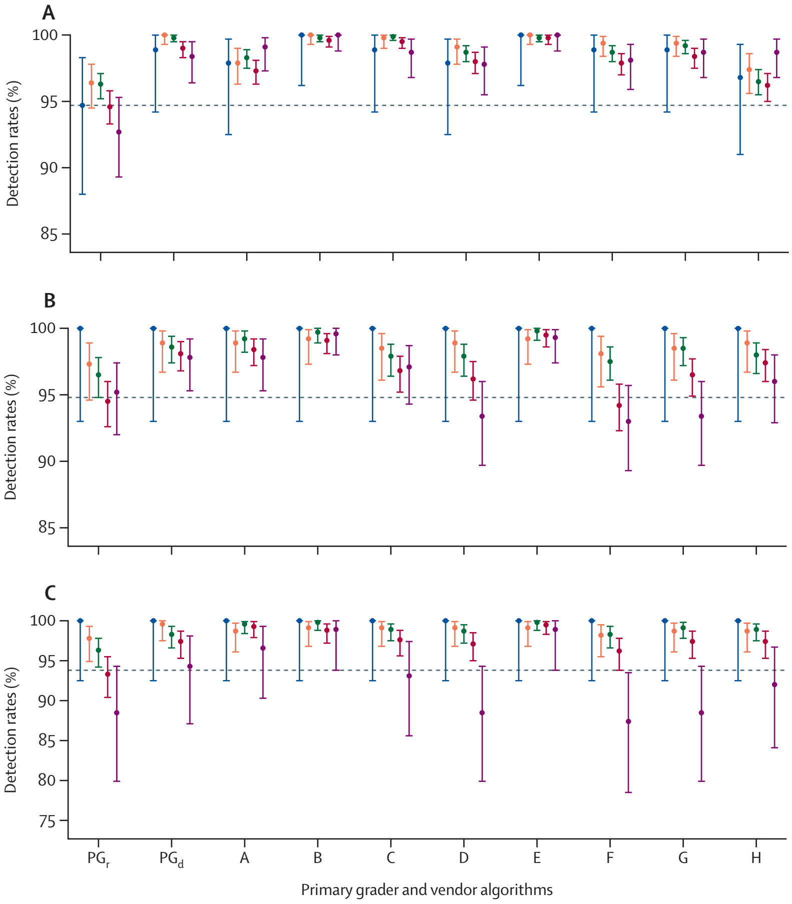
Inclusion of the 6/12 visual acuity criterion resulted in increased detection of referable diabetic retinopathy (from 87% to 99% across vendors; appendix p 17). This increase did not materially affect the already high sensitivity for R2, R3, and R3A, but improved sensitivity for R1M1 (83·6% to 98·7%; appendix pp 13, 19, 28–35). Overall screen positive rates slightly increased (35% to 77%) but positive and negative predictive values were not materially altered (appendix p 18).

Six of the eight algorithms had a second operating threshold (designed to detect more severe diabetic retinopathy), which had lower false positive and detection rates for referable diabetic retinopathy, but sensitivities for R2 and R3 remained high for some algorithms (appendix pp 36–47).

Among 18 887 cases with referable diabetic retinopathy as per reference standard with good acuity that were ARIAS test-negative, between five and 24 (0·03–0·13%) were urgent referrals to ophthalmology for diabetic retinopathy, and another six to 56 (0·03–0·3%) were referred to ophthalmology departments for diabetic retinopathy (appendix p 52). Most cases were reviewed in a local imaging pathway at 1, 3, or 6 months.

Discussion

Eight CE-marked commercial ARIAS did as well as or better than primary human graders for the detection of moderate-to-severe non-proliferative or proliferative diabetic retinopathy (with or without referable maculopathy), and did not show systematic differences (bias) across subgroups of ethnicity, sex, and IMD quintiles, but had lower performance (in agreement with human graders) for the



Age group (years)	Primary grader classification (95% CI)	
	% referable diabetic retinopathy	% diabetic retinopathy present
	R2M0 + R2M1	
<30	94·7 (88·0–98·3)	98·9 (94·2–100)
30 to <45	96·4 (94·5–97·8)	100 (99·3–100)
45 to <60	96·3 (95·2–97·1)	99·8 (99·5–100)
60 to <75	94·6 (93·3–95·8)	99·0 (98·3–99·5)
≥75	92·7 (89·3–95·3)	98·4 (96·4–99·5)
	R3M0 + R3M1	
<30	100·0 (93·0–100·0)	100 (93·0–100)
30 to <45	97·3 (94·6–98·9)	98·9 (96·7–99·8)
45 to <60	96·5 (94·8–97·8)	98·6 (97·4–99·4)
60 to <75	94·5 (92·6–96·0)	98·1 (96·8–99)
≥75	95·2 (92·0–97·4)	97·8 (95·3–99·2)
	R3AM0 + R3AM1	
<30	100·0 (92·5–100·0)	100 (92·5–100)
30 to <45	97·8 (94·9–99·3)	99·6 (97·5–100)
45 to <60	96·3 (94·2–97·8)	98·3 (96·6–99·3)
60 to <75	93·3 (90·4–95·5)	97·4 (95·3–98·7)
≥75	88·5 (79·9–94·3)	94·3 (87·1–98·1)

***Figure 3:* Performance for PGs and ARIAS (vendor algorithms A–H) by age group**

Numerical values represent the proportion of encounters with specified reference standard diabetic retinopathy grade that were classified as referable diabetic retinopathy (PG_r) or diabetic retinopathy present (PG_d) by PG. For vendor algorithms, A to H values represent the proportion of encounters that were ARIAS test positive. Error bars are 95% CIs. (A) Reference standard diabetic retinopathy grade R2; moderate-to-severe non-proliferative diabetic retinopathy. (B) Reference standard diabetic retinopathy grade R3; proliferative diabetic retinopathy including those with or without diabetic maculopathy (grades M0 and M1). (C) Reference standard diabetic retinopathy grade R3A; active proliferative diabetic retinopathy, including those with or without diabetic maculopathy (grades M0 and M1). Blue indicates <30 years; orange indicates 30 years to <45 years; green indicates 45 years to <60 years; red indicates 60 years to <75 years; and purple indicates ≥75 years. The horizontal dashed line is the lower 95% CI limit for PG_r for all subgroups combined. ARIAS=automated retinal image analysis systems. M0=No observable diabetic maculopathy. M1=diabetic maculopathy. PG=primary grader. R2=moderate-to-severe non-proliferative (pre-proliferative) retinopathy. R3=proliferative retinopathy.

oldest age group (likely due to poor image quality in older people, eg, with cataracts or small pupils hindering diabetic retinopathy detection by both humans and ARIAS). ARIAS could reduce the need for human grading by 26–77%. To our knowledge, this is the largest open-label, vendor-independent, head-to-head evaluation of eight ARIAS in the same computational environment, including over 200 000 encounters, and around 1·2 million images with nearly 25 000 cases of referable diabetic retinopathy from a routine, high-volume national diabetic eye screening programme. There is good representation across population subgroups, including ethnicity, age, level of deprivation, and the spectrum of diabetic eye disease.

We showed that NEL DESP primary graders' output used as a comparator for ARIAS in this study mapped well with another large screening centre (SEL DESP) for active proliferative diabetic retinopathy, and was unlikely to have missed cases of R3. Hence, we are confident that our primary graders are representative of primary graders in other quality-assured screening centres.[12]

Previous standards based on expert opinion suggested detection rates for any diabetic retinopathy should be at least 90%.[27] The current quality assurance standard for human graders in the English NHS DESP[25] for referrable diabetic retinopathy is 85%. If ARIAS were to follow the same quality assurance standards as set for primary human graders, the minimum acceptable sensitivity for referable diabetic retinopathy should be at least 85%. Although several ARIAS achieved this standard, there are differences in how they perform. From a safety perspective, we propose that the lower limit of the 95% CI for R2, R3, and R3A sensitivity (regardless of any statistical test of significance) be considered alongside quality assurance standards for primary graders by population subgroups (eg, age, ethnicity, and visual acuity) relevant to the intended screening setting.[16]

Agreement between human graders is known to be heterogeneous, especially for less severe diabetic retinopathy grades (R1) or milder forms of maculopathy.[9,28,29] Human graders miss 11% of cases with R1 and 2–4% of cases with referable diabetic retinopathy,[28] likely contributing to the observed variation in ARIAS sensitivity across and within vendors for diabetic retinopathy grades R1 and R1M1. Inclusion of a visual acuity criterion increased sensitivity overall, especially for grades R1 and R1M1, with a small associated rise in screen positive rates. Among those with reference standard referable diabetic retinopathy and good acuity, ARIAS missed less than 0·5% of cases that were subsequently referred for diabetic retinopathy or other conditions. We and others recommend ARIAS as a first pass to triage patients into low risk (no human grading) and medium risk or high risk (receive human grading);[9–12] hence lower specificity (ie, higher false positive rate) could still be cost-effective by increasing grading capacity,[12] even with a visual acuity criterion. Differences in ARIAS performance across factors such as age might necessitate tailored ARIAS to ensure equity. Our previous work estimated that annual screening of around 2 million people with ARIAS to triage for human grading could save the NHS £8–10 million per year, with earlier versions of ARIAS having lower test performance.[12] Updated cost-effectiveness analyses are needed to compare current practice with triage afforded by different ARIAS and comparative analyses of annual or biennial recall for low-risk groups.

Most published ARIAS performance comparisons are weakened by single ARIAS evaluations in populations with low sociodemographic diversity,[30] small sample sizes, unspecified pre-selection or pre-processing of retinal images, and unspecified image capture systems, grading protocols, and reference standards.[9,30] Our approach circumvents these issues by design. The strengths of our study included multiple ARIAS head-to-head comparisons on a large, appropriately powered, sociodemographically diverse, clinically relevant dataset using the same computational environment and executed by a vendor-neutral research team. Although based at one screening centre, this centre is one of the largest NHS screening centres, serving one of the most diverse and deprived populations in the country. Validation of our primary graders, along with our previous work,[10] supports the generalisability of the findings. Although the reference standard (final human grade) is not an absolute ground truth and might have classification bias, as not all encounters received higher-level grading, the reference standard provided a pragmatic solution for real-world evaluation. The eight self-selected vendors might represent algorithms developed on populations similar to the current study, potentially leading to better performance. Other vendors might have declined to participate due to concerns about poor results. These factors underscore the importance of conducting such evaluations with prespecified standards and analysis plan, akin to phase 3 clinical trials for regulatory approval and clinical application of pharmaceutical agents.

Our approach aligns with a governmental review of equity in medical devices,[15] highlighting the need to assess algorithmic fairness and equity using real-world data before deployment in health care.[13–15,17] We have developed a transferable framework for the evaluation of clinical AI,[16] ensuring algorithms meet predefined standards for fairness and trustworthiness before being commissioned. By focusing on algorithmic fairness, we aim to promote equal opportunities for ARIAS in health-care services, preventing monopolies and encouraging investment. Our public engagement efforts[31] aim to build trust, innovation, and cost-effective advancements.

In conclusion, we compared multiple ARIAS across diverse populations and found that those matching or exceeding human grader sensitivity for moderate-to-severe non-proliferative (R2) or proliferative (R3) diabetic retinopathy—with or without diabetic maculopathy—can safely be used in global screening programmes to triage and could reduce human grading by up to 80%.

The Artificial Intelligence Automated Retinal Image Analysis Systems (ARIAS) Research Group

John Anderson (Diabetes and Endocrinology, Homerton Healthcare NHS Foundation Trust, London, UK), Sarah Barman (School of Computer Science and Mathematics, Kingston University, London, UK), Louis Bolter (Diabetes and Endocrinology, Homerton Healthcare NHS Foundation Trust, London, UK), Ryan Chambers (Diabetes and Endocrinology, Homerton Healthcare NHS Foundation Trust, London, UK), Lakshmi Chandrasekaran (St George's School of Health and Medical Sciences, City St George's, University of London, London, UK), Umar Chaudhry (St George's School of Health and Medical Sciences, City St George's, University of London, London, UK), Catherine Egan (NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, UK), Jiri Fajtl (School of Computer Science and Mathematics, Kingston University, London, UK), Aaron Lee (Department of Ophthalmology, University of Washington, Seattle, WA, USA), Samantha Mann (South East London Diabetic Eye Screening Programme, Guy's and St Thomas' NHS Foundation Trust, London, UK), Abraham Olvera-Barrios (NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, UK), Christopher G Owen (St George's School of Health and Medical Sciences, City St George's, University of London, London, UK), Paolo Remagnino (Department of Computer Science, Durham University, Durham, UK), Alicja R Rudnicka (St George's School of Health and Medical Sciences, City St George's, University of London, London, UK), Adnan Tufail (NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, UK), Charlotte Wahlich (St George's School of Health and Medical Sciences, City St George's, University of London, London, UK), Laura Webster (South East London Diabetic Eye Screening Programme, Guy's and St Thomas' NHS Foundation Trust, London, UK), Roshan Welikala (School of Computer Science and Mathematics, Kingston University, London, UK), Kathryn Willis (St George's School of Health and Medical Sciences, City St George's, University of London, London, UK)

Contributors

All authors contributed to this manuscript. ARR, CE, AT, JA, CGO, SAB, PR, and AL conceived and designed the study and raised funding. ARR, RS,

CGO, RAW, SAB, RC, JA, LB, JF, AO-B, SM, and LW collected data for the study and undertook data curation and data management. ARR, JF, and SB undertook project administration. ARR, RS, RAW, JF, SAB, JA, and RC verified the raw data, had access to the data for analyses of the images, and had access to the diabetic eye screening programme data linked to image analyses outcomes. ARR, CE, AT, RC, LB, JA, LW, SM, CGO, SAB, and AO-B supervised the data collection and validation. ARR wrote the first draft of the report, which all authors contributed to. All authors critically appraised the text and figures. ARR, RS, and CGO had direct access to and verified the data reported in the manuscript. ARR and CGO are responsible for data integrity and will act as guarantors. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication. The ARIAS Research Group all read, fed back on, and approved the final manuscript.

Declaration of interests

AT has received consulting fees from AbbVie, Adverum, Annexon, Apellis Pharmaceuticals, Aviceda Therapeutics, Boehringer Ingelheim, EyePoint Pharmaceuticals, Genentech-Roche, Iveric Bio, Janssen Pharmaceuticals, Nanoscope, Novartis, Ocular Therapeutix, Opthea, Oxurion, Regenxbio, and Thea Pharma; honoraria from Ocular Therapeutix, Opthea, and Boehringer Ingelheim; and has stocks or shares from Celesio and Oculogics. CE has received fees from Heidelberg Engineering, Inozyme Pharma, and Boehringer Ingelheim and reports travel costs and receipt of equipment from Lowy Medical Research Institute. SM is president of the British Association of Retinal Screeners. AL reports grants or contracts from the National Eye Institute (K23EY029246), National Institutes of Health (OT2OD032644), Research to Prevent Blindness, Latham Vision Research Innovation Award, C Dan and Irene Hunter Endowed Professorship, Santen, Topcon, Carl Zeiss Meditec, Amazon, and Meta; consulting fees from the US Food and Drug Administration, Genentech, Sanofi, Johnson & Johnson, Gyroscope, Boehringer Ingelheim, Roche, and Janssen; honoraria from Topcon and Alcon (including travel costs); and receipt of equipment from iCare World, Optomed, Heidelberg, Microsoft, and Regeneron. All other authors declare no competing interests.

Data sharing

In compliance with our data protection impact assessment, sharing of individual patient data is not possible and all data are stored by the data controller, the Homerton Healthcare NHS Foundation Trust. A fully anonymised dataset is available from the North East London Diabetic Eye Screening Programme upon reasonable request to JA.

Acknowledgments

We thank Niall Canavan, Sean Devine, and Asif Mirza from the Homerton Healthcare NHS Foundation Trust IT department for their continued support throughout this project. We thank our study advisory group members—Rosalind Given-Wilson, Alastair Denniston, Kevin Dunbar, Samantha Mann, and Fiona Martin—for their valuable contributions to this work. Their insights, feedback, and ongoing support throughout the study have been instrumental in shaping the research and ensuring its relevance and rigor. This work was funded by the NHS Transformation Directorate and The Health Foundation and managed by the National Institute for Health and Social Care Research (Ethnic differences in performance and perceptions of Artificial Intelligence retinal image analysis systems for the detection of diabetic retinopathy in the NHS Diabetic Eye Screening Programme https://fundingawards.nihr.ac.uk/award/AI_HI200008) and the Wellcome Trust Collaborative Award (224390/Z/21/Z). The views expressed in this publication are those of the authors and not necessarily those of the NHS Transformation Directorate, The Health Foundation, National Institute for Health and Care Research, or the UK Department of Health and Social Care. AT and CE received a proportion of their financial support from the UK Department of Health through an award made by the National Institute for Health Research to Moorfields Eye Hospital NHS Foundation Trust and University College London Institute of Ophthalmology for a Biomedical Research Centre for Ophthalmology.

References

1 Sun H, Saeedi P, Karuranga S, et al. IDF Diabetes Atlas: global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res Clin Pract* 2022; **183:** 109119.

2 Hex N, Bartlett C, Wright D, Taylor M, Varley D. Estimating the current and future costs of type 1 and type 2 diabetes in the UK, including direct health costs and indirect societal and productivity costs. *Diabet Med* 2012; **29:** 855–62.

3 Bellemo V, Lim G, Rim TH, et al. Artificial intelligence screening for diabetic retinopathy: the real-world emerging application. *Curr Diab Rep* 2019; **19:** 72.

4 Liew G, Michaelides M, Bunce C. A comparison of the causes of blindness certifications in England and Wales in working age adults (16–64 years), 1999–2000 with 2009–2010. *BMJ Open* 2014; **4:** e004015.

5 Public Health England. NHS screening programmes in England. 1 April 2017 to 31 March 2018. 2019. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/783537/NHS_Screening_Programmes_in_England_2017_to_2018_final.pdf (accessed Feb 16, 2023).

6 Office for Health Improvement & Disparities. Diabetic eye screening pathway requirements specification. October, 2023. https://www.gov.uk/government/publications/diabetic-eye-screening-pathway-requirements-specification/diabetic-eye-screening-pathway-requirements-specification (accessed Oct 21, 2023).

7 Scanlon PH. The contribution of the English NHS Diabetic Eye Screening Programme to reductions in diabetes-related blindness, comparisons within Europe, and future challenges. *Acta Diabetol* 2021; **58:** 521–30.

8 Tufail A, Rudisill C, Egan C, et al. Automated diabetic retinopathy image assessment software: diagnostic accuracy and cost-effectiveness compared with human graders. *Ophthalmology* 2017; **124:** 343–51.

9 Zhelev Z, Peters J, Rogers M, et al. Automated grading in the Diabetic Eye Screening Programme. External review against programme appraisal criteria for the UK National Screening Committee. Nov 24, 2021. https://assets.publishing.service.gov.uk/media/619e80088fa8f5037e8ccb2f/Evidence_summary_AI_in_DESP_2021.pdf (accessed Feb 16, 2023).

10 Heydon P, Egan C, Bolter L, et al. Prospective evaluation of an artificial intelligence-enabled algorithm for automated diabetic retinopathy screening of 30 000 patients. *Br J Ophthalmol* 2021; **105:** 723–28.

11 Bhaskaranand M, Ramachandra C, Bhat S, et al. The value of automated diabetic retinopathy screening with the EyeArt System: a study of more than 100 000 consecutive encounters from people with diabetes. *Diabetes Technol Ther* 2019; **21:** 635–43.

12 Tufail A, Kapetanakis VV, Salas-Vega S, et al. An observational study to assess if automated diabetic retinopathy image assessment software can replace one or more steps of manual imaging grading and to determine their cost-effectiveness. *Health Technol Assess* 2016; **20:** 1–72.

13 Noor P. Can we trust AI not to further embed racial bias and prejudice? *BMJ* 2020; **368:** m363.

14 Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019; **366:** 447–53.

15 UK Department of Health and Social Care. Equity in medical devices: independent review—final report. March 11, 2024. https://www.gov.uk/government/publications/equity-in-medical-devices-independent-review-final-report (accessed March 20, 2024).

16 Fajtl J, Welikala RA, Barman S, et al. Trustworthy evaluation of clinical AI for analysis of medical images in diverse populations. *NEJM AI* 2024; **1:** AIoa2400353.

17 Sujan M, Smith-Frazer C, Malamateniou C, et al. Validation framework for the use of AI in healthcare: overview of the new British standard BS30440. *BMJ Health Care Inform* 2023; **30:** e100749.

18 Early Treatment Diabetic Retinopathy Study Research Group. Fundus photographic risk factors for progression of diabetic retinopathy. ETDRS report number 12. *Ophthalmology* 1991; **98** (suppl)**:** 823–33.

19 Wilkinson CP, Ferris FL 3rd, Klein RE, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology* 2003; **110:** 1677–82.

20 Olvera-Barrios A, Owen CG, Anderson J, et al, and the ARIAS Research Group. Ethnic disparities in progression rates for sight-threatening diabetic retinopathy in diabetic eye screening: a population-based retrospective cohort study. *BMJ Open Diabetes Res Care* 2023; **11:** e003683.

21 NHS England. Diabetic eye screening: professional guidance. June 22, 2023. https://www.gov.uk/government/collections/diabetic-eye-screening-commission-and-provide (accessed June 30, 2023).

22 Geographic Data Service. Index of Multiple Deprivation (IMD). 2019. https://data.cdrc.ac.uk/dataset/index-multiple-deprivation-imd (accessed Aug 30, 2024).

23 NHS England. Diabetic eye screening pathways: patient, grading, referral, surveillance. Sept 27, 2024. https://www.gov.uk/government/publications/diabetic-eye-screening-pathways-patient-grading-referral-surveillance (accessed Oct 30, 2024).

24 NHS England. Diabetic eye screening: retinal image grading criteria. https://www.gov.uk/government/publications/diabetic-eye-screening-retinal-image-grading-criteria (accessed Oct 30, 2024).

25 NHS. The management of grading quality. Good practice in the quality assurance of grading. March 22, 2016. https://assets.publishing.service.gov.uk/media/5a80521b40f0b62305b8a76c/The_Management_of_Grading.pdf (accessed Nov 23, 2023).

26 NHS England. Diabetic eye screening: programme specific operating model. https://www.gov.uk/government/publications/diabetic-eye-screening-internal-and-external-quality-assurance/diabetic-eye-screening-programme-specific-operating-model (accessed Nov 3, 2024).

27 Taylor R, Broadbent DM, Greenwood R, Hepburn D, Owens DR, Simpson H. Mobile retinal screening in Britain. *Diabet Med* 1998; **15:** 344–47.

28 Scanlon PH, Aldington SJ, Leal J, et al. Development of a cost-effectiveness model for optimisation of the screening interval in diabetic retinopathy screening. *Health Technol Assess* 2015; **19:** 1–116.

29 Teoh CS, Wong KH, Xiao D, et al. Variability in grading diabetic retinopathy using retinal photography and its comparison with an automated deep learning diabetic retinopathy screening software. *Healthcare (Basel)* 2023; **11:** 1697.

30 Rajesh AE, Davidson OQ, Lee CS, Lee AY. Artificial intelligence and diabetic retinopathy: AI framework, prospective studies, head-to-head validation, and cost-effectiveness. *Diabetes Care* 2023; **46:** 1728–39.

31 Willis K, Chaudhry UAR, Chandrasekaran L, et al. What are the perceptions and concerns of people living with diabetes and National Health Service staff around the potential implementation of AI-assisted screening for diabetic eye disease? Development and validation of a survey for use in a secondary care screening setting. *BMJ Open* 2023; **13:** e075558.