AMERICAN ACADEMY OF OPHTHALMOLOGY®

A Comparison of Machine Learning and Human Graders for Glaucoma Diagnosis from Fundus Images for Population Screening

Thomas R.P. Taylor, BMBCh,[1,] Justin Khasentino, MS,[2,*] Robert N. Luben, PhD,[1,*] Laura A. Meliante, MD,[1,3] Kelsey V. Stuart, MBBCh, PhD,[1] Dun Jack Fu, BMBCh, PhD,[1] Michelle P.Y. Chan, PhD, FRCOphth,[1] David C. Broadway, MD, FRCOphth,[4] Farhad Hormozdiari, PhD,[7] Andrew Carroll, PhD,[2] Ruiqi Hu, MSc,[1] Wessel Verburg, MD,[1] Mahantesh I. Biradar, PhD,[1] Pearse A. Keane, MD, FRCOphth,[1] Paul J. Foster, PhD, FRCS(Ed),[1,5,6] Cory Y. McLean, PhD,[7,†] Anthony P. Khawaja, PhD, FRCOphth[1,†]*

Purpose: To compare the accuracy of vertical cup-disc ratios (VCDRs), ascertained by machine learning (ML) versus human graders, from fundus images for glaucoma detection. This study uses population-based data, with a disease prevalence and case mix that is more reflective of an unselected patient population than conventional case–control studies, with the aim of developing improved glaucoma screening tests.

Design: Cross-sectional analysis of a population-based study.

Participants: A total of 6304 participants of the European Prospective Investigation of Cancer (EPIC)-Norfolk Eye Study with color fundus images gradable by humans and ML in both eyes.

Methods: Vertical cup-disc ratio was independently estimated from 2-dimensional fundus images of EPIC-Norfolk Eye Study participants by trained human graders (human-derived VCDR [H-VCDR]) and an externally trained, open-access ML model (ML-derived VCDR [ML-VCDR]). A neural network trained on 81 830 ophthalmologist-labeled images was used to generate pseudo-labels for over 100 000 UK Biobank images, on which ML-VCDR was subsequently trained. Glaucoma status was ascertained by tertiary-center specialist examination. Predictive performance of VCDR for glaucoma status was examined using logistic regression. Machine learning–derived VCDR estimates were additionally compared with a popular open-source ML model (AutoMorph) and scanning laser ophthalmoscopy (Heidelberg Retinal Tomography [HRT]).

Main Outcome Measures: Area under the receiver operating characteristic curve (AUROC) explained variance (McFadden's pseudo-R^2).

Results: Of 6304 participants (mean age, 68 years; 57% women), 696 had glaucoma or suspect status in at least 1 eye. For left eyes, H-VCDR and ML-VCDR explained 17% (95% confidence interval [CI], 14.7–20.4) and 31% (95% CI, 27.9–33.9) of glaucoma status variance and had an AUROC of 79% (95% CI, 76.8–81.2) and 88% (95% CI, 86.3–89.1), respectively. Right eye H-VCDR and ML-VCDR explained 20% (95% CI, 16.9–22.6) and 35% (95% CI, 32.4–37.9) of the variance and had an AUROC of 81% (95% CI, 78.6–82.5) and 90% (95% CI, 88.7–91.0), respectively. Machine learning–derived VCDR also performed better than AutoMorph and HRT at predicting glaucoma status from VCDR estimations.

Conclusions: In this population-based setting, ML far outperformed trained human graders at predicting specialist-ascertained glaucoma status from fundus images. This provides promise for ML-supported strategies for glaucoma population screening.

Financial Disclosure(s): Proprietary or commercial disclosure may be found in the Footnotes and Disclosures at the end of this article. *Ophthalmology 2026;133:1447-1457* *© 2026 American Academy of Ophthalmology, Inc. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).*

Supplemental material available at www.aaojournal.org.

See Commentary on page 1458

Glaucoma is the leading cause of irreversible blindness worldwide.[1] Despite a range of treatment options, including medications, laser, and surgery, people continue to go blind from the disease. A significant contributing factor to this are those who present with late-stage disease are unaware that they had a sight-threatening condition until they are suffering from advanced visual field loss.[2,3] A recent meta-analysis indicated that 56% of people living with glaucoma are currently undiagnosed.[4] Lack of attendance for routine eye examinations, particularly among socioeconomically deprived groups,[3,5] continues to contribute to later presentation and inequities in glaucoma diagnosis.

© 2026 American Academy of Ophthalmology, Inc
Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
http://doi.org/10.1016/j.ophtha.2026.07.043
ISSN 0161-6420/26

Population-based screening has been proposed as a means to address this gap. However, governmental screening committees currently recommend against glaucoma screening because of concerns regarding the accuracy of available tests in low-prevalence settings.[6,7] Proposed screening strategies, including limited testing by technicians and comprehensive assessment by specialist optometrists, are likely to be cost-ineffective in the general population.[8,9] Excessive false-positive referrals would be unsustainable for already overburdened health care services. Fundus imaging is increasingly accessible, and machine learning (ML)—based analysis of fundus images has emerged as a potential approach to improve diagnostic accuracy through automated detection of signs of glaucoma such as enlarged vertical cup-disc ratio (VCDR), notching, and disc hemorrhage. Reports of high sensitivities and specificities for case detection, however, are predominantly derived from case—control studies that exclude diagnostically ambiguous cases and do not reflect the heterogeneity or prevalence of screening populations.[10–12] Far fewer studies assess ML performance in population data sets, and those that do typically report poorer results.[13,14]

Our study seeks to address this evidence gap. We evaluated how an ML model for VCDR estimation (ML-derived VCDR [ML-VCDR]), trained on a large number of expert-labeled fundus images, performs as a test for glaucoma screening in an external, population-based data set. We compared this performance with independent, human-graded, VCDR estimates of the same images. This provides further evidence regarding ML-based glaucoma detection in a population with a disease prevalence and case mix that is more representative of a screening scenario and directly compares ML-VCDR with human assessment. We have made ML-VCDR openly available for use by other researchers.

Methods

The European Prospective Investigation of Cancer-Norfolk Eye Study

We analyzed data from the European Prospective Investigation of Cancer (EPIC)-Norfolk Eye Study, a cross-sectional population-based eye study of 8623 participants of a larger, ongoing prospective cohort study.[15] European Prospective Investigation of Cancer-Norfolk was a UK center for the 10-country collaborative EPIC study and recruited 30 445 participants aged 40 to 79 years from 35 participating general practices in East Anglia between 1993 and 1997.[16]

All participants who had not died or opted out of the study were invited to take part in the EPIC-Norfolk Eye Study between 2004 and 2011. Undilated ocular investigations were carried out by trained nurses. Detailed methods have previously been published.[17] Macula-centered nonstereo fundus photographs including the optic disc were taken using a TRC-NW6S nonmydriatic retinal camera and IMAGEnet Telemedicine System (Topcon Corporation). Optic nerve head anatomy was assessed using scanning laser ophthalmoscopy (SLO) (Heidelberg Retinal Tomography [HRT] II; Heidelberg Engineering), with only images of acceptable quality included (topography standard deviation $\leq$40 μm). Refractive error was measured using a Humphrey Auto-Refractor 500 (Humphrey Instruments).

The ascertainment of glaucoma status in EPIC-Norfolk has also been previously reported.[18] All participants underwent visual acuity and intraocular pressure testing combined with scanning laser polarimetry, SLO, and fundus photography. A 24-2 central threshold visual field test was also performed on 1 in 10 participants with otherwise normal findings, as well as on those with abnormal findings on scanning laser polarimetry or SLO. Those who met the predefined referral criteria (Table S1, available at www.aaojournal.org) were referred to a tertiary eye center for definitive assessment by a consultant ophthalmologist specializing in glaucoma. Glaucoma was defined as the presence of characteristic structural abnormalities of the optic disc and visual field loss, with no other explanations for the disc and field appearances. Suspected glaucoma was defined as the presence of early or minor glaucomatous disc features, associated with a normal visual field or the absence of visual field data. All glaucoma suspects were deemed to be at sufficient risk of vision loss to require ongoing monitoring in the specialist center. Therefore, we defined glaucoma status for this study as participants meeting the case or suspect definitions.

Fundus images were graded for VCDR and other disc features as well as photo quality in the Reading Centre at Moorfields Eye Hospital by human graders who had undertaken accredited training relevant to individual diseases, including glaucoma. Disc measurements were made directly on screen along the vertical axis of the optic disc using Adobe Photoshop CS5 software (Adobe, 2010). The graders were masked to the ophthalmic history of the participants and other investigation results. Moorfields Eye Hospital Reading Centre quality assurance procedures were used, including intraobserver and interobserver evaluation of grading. A total of 6304 participants with color fundus images gradable by both humans and ML-VCDR in both eyes were included in the main analysis. The process of participant selection is illustrated in Figure 1.

Ethical Approval

All participants provided written informed consent for the EPIC-Norfolk Study. The study adhered to the tenets of the Declaration of Helsinki. It was approved by the Norwich Local Research Ethics Committee (05/Q0101/191) and the East Norfolk & Waveney NHS Research Governance Committee (2005EC07L).

ML-VCDR Model Development

The process of the ML-VCDR model development is illustrated in Figure S1 (available at www.aaojournal.org). The starting point for the model used in this analysis was a previously published open-code, closed-weights neural network trained on 81 830 non-UK Biobank (UKB), ophthalmologist-labeled fundus images. These images were assessed for glaucomatous optic nerve head features including focal notching, nerve fiber layer defects, and disc hemorrhages and were given a grade for referable glaucomatous optic neuropathy.[19] This "teacher" model, trained on large, expert-labeled data, was used to generate pseudo-labels of VCDR, image gradeability, glaucoma referral risk, and eye laterality for 175 337 UKB images. These pseudo-labels were previously shown to correlate strongly with expert-assigned and adjudicated labels on a subset of UKB images ($N = 2115$, $R = 0.89$).[19] It was not possible to run the "teacher" model directly on EPIC-Norfolk data because of data protection requirements between parties. To create a fully open-source, open-weights model for this work, we created the ML-VCDR model by using these published pseudo-labels for supervised training. Data were randomly split into training (80%; 140 056 images) and validation (20%; 34 684 images) sets at the

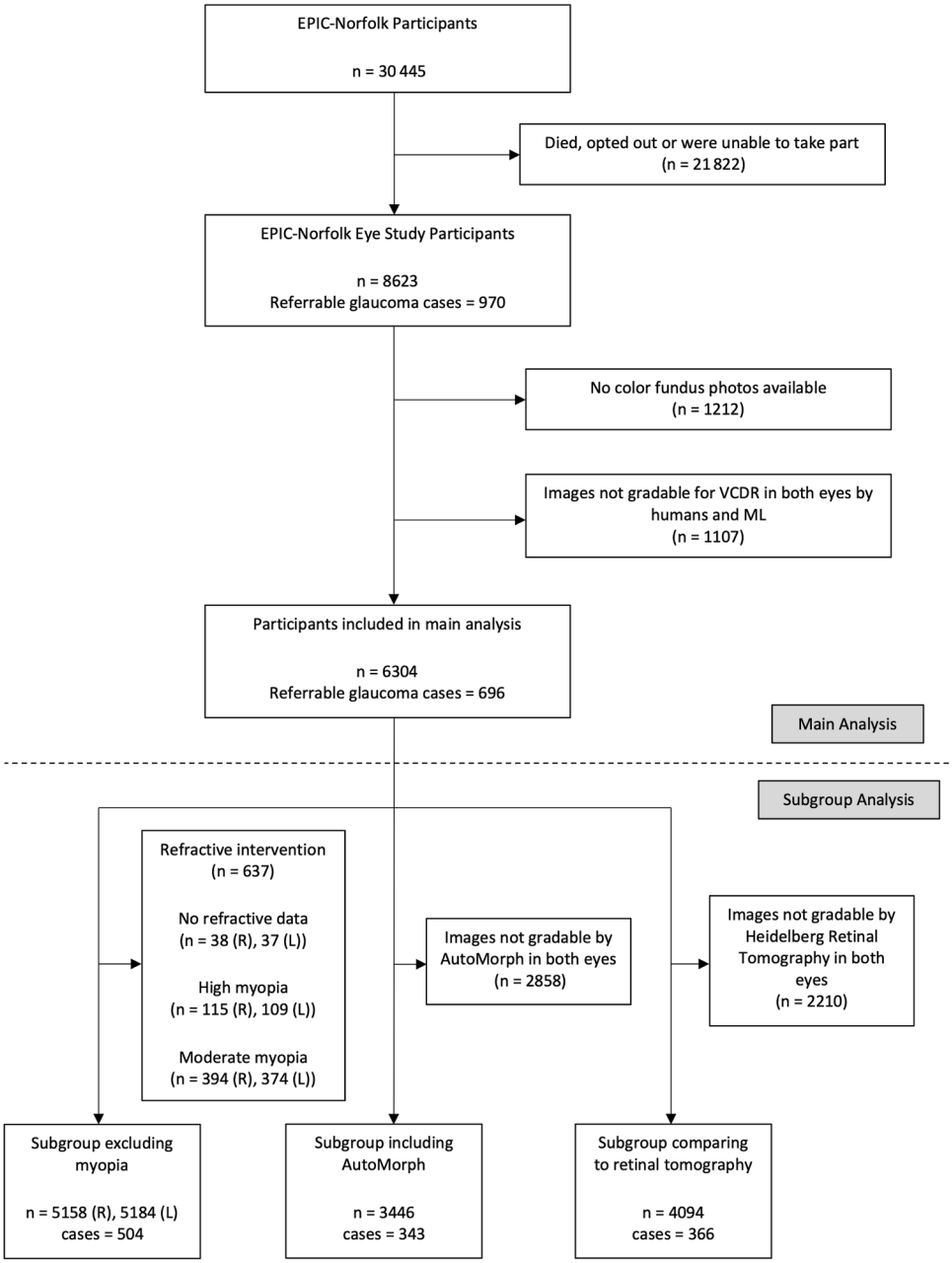


Figure 1. Flowchart outlining participant selection for main and subgroup analyses.

patient level, dropping individuals to ensure no related individuals spanned splits. Images were cropped to 1904 × 1904 pixels to remove metadata and corner text. The ML-VCDR model is a multitask deep convolutional neural network based on the Inception v3 architecture[20] that uses the same architecture, hyperparameters, and training recipe as the "teacher" model.[19] The convolutional neural network backbone was warm-started from a checkpoint pretrained on the ImageNet data set.[21] During training, image augmentation and early stopping were applied. The final model checkpoint was selected to minimize the VCDR mean squared error in the evaluation set. Images were defined as gradable if their predicted gradeability exceeded 0.7, as previously validated.[19] Machine learning–derived VCDR was trained de novo using the pseudo-labels as supervisory targets, with the aim of enabling phenotyping of fundus images in large biobank or population data sets. Machine learning–derived estimates of VCDR, image quality, referability risk, and laterality were used in the analysis. The ML-VCDR code and weights are openly available online (https://github.com/Google-Health/genomics-research/tree/main/ml-based-vcdr-ext).

Subgroup Analysis Excluding Moderate and High Myopia

A subgroup analysis was performed to determine the effect on results after participants with moderate and high myopia were excluded because of the potentially increased difficulty in

assessing the optic disc in myopic patients. As illustrated in Figure 1, 637 participants were excluded because they had an intervention that may have affected their refractive status (cataract surgery, refractive surgery, and previous retinal detachment). For right eyes (OD), 115 participants with high myopia (average spherical equivalent worse than or equal to −6.00) and 394 with moderate myopia (average spherical equivalent between −3.00 and −6.00) were then removed from the analysis. For left eyes (OS), 109 with high myopia and 374 with moderate myopia were removed.

AutoMorph

Separate VCDR estimates were also obtained by running AutoMorph on the EPIC-Norfolk images.[22] AutoMorph is a deep learning pipeline for automated analysis of retinal vascular morphology and optic disc/cup segmentation (https://github.com/rmaphoh/AutoMorph). AutoMorph is a highly cited and open-access pipeline for fundus image analysis and was therefore selected as a comparator for ML-based VCDR assessment. AutoMorph features its own module for image quality grading, and poor-quality images are filtered out before morphological feature measurement.

Statistical Analysis

Statistical analysis was conducted in RStudio (version 2025.05.1+513; Posit Software, PBC) using the pscl, pROC, and boot packages. Graphs were created using ggplot2. Logistic regression models were fitted for each method of VCDR estimation, with VCDR as the predictor and glaucoma status (any glaucoma or suspect-glaucoma case, analyzed on a per-eye basis) as the dependent variable. Explained variance was quantified using McFadden's pseudo-R^2. Receiver operating characteristic (ROC) curves were generated, and the area under the ROC curve (AUROC) was calculated to assess discriminatory performance. Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were evaluated at the following 2 operating points: (1) the threshold that maximized the Youden index and (2) a threshold corresponding to a fixed specificity of 95%, identified from the ROC curve.

Uncertainty in performance metrics was estimated using nonparametric bootstrap resampling with 1000 replicates. Models and ROC curves were refitted within each bootstrap resample, with point estimates calculated from the full data set and 95% confidence intervals (CIs) derived from the bootstrap distributions.

Direct comparisons between methods of deriving VCDR (and between ML-VCDR and ML-derived glaucoma risk score) were performed using paired bootstrap resampling. Identical bootstrap samples of eyes were used to compute all performance metrics for both methods within each resample. For each metric and operating point, paired differences were calculated across bootstrap replicates. Statistical significance of between-method differences was assessed using 95% percentile CIs derived from the paired differences and was considered statistically significant if the CI did not include zero.

Results

Participant Characteristics

Of the 8623 participants with the EPIC-Norfolk Eye Study, 7411 participants had color fundus images taken of at least 1 eye. A total of 6304 participants had gradable images by humans and the ML-VCDR model in both eyes. The mean age in this group was 67.8 years (range, 48.4−91.9); 56.6% were women. Approximately 99.4% of participants identified as White. A total of 696 participants had glaucoma in at least 1 eye or were considered suspects of sufficient risk to warrant specialist monitoring (582 OD and 576 OS). The process of participant inclusion is shown in Figure 1.

ML-VCDR vs. Human Graders (human-derived VCDR)

The distribution of VCDR estimates for ML and human graders (human-derived VCDR [H-VCDR]) is illustrated in Figure 2A. Machine learning−derived VCDR had a lower median estimate of 0.35 for both OD and OS compared with H-VCDR median estimates of 0.39 for both eyes. Machine learning−derived VCDR had a wider interquartile range (IQR) (OD: IQR = 0.21; OS: IQR = 0.20) than H-VCDR (OD and OS: IQR = 0.11); however, it estimated a higher percentage of VCDRs to be >0.8 (0.80% for both eyes for ML-VCDR, compared with 0.05% [OD] and 0.03% [OS] for H-VCDR).

Figure 3A, B illustrates the association between glaucoma status and VCDR as estimated by humans and ML. Right eye ML-VCDR and H-VCDR explained 35% (95% CI, 32.4−37.9) and 20% (95% CI, 16.9−22.6) of glaucoma status variance, respectively. For OS, ML-VCDR and H-VCDR explained 31% (95% CI, 27.9−33.9) and 17% (95% CI, 14.7−20.4) of the variance of glaucoma status, respectively. ML-VCDR had an AUROC for glaucoma status of 90% (95% CI, 88.7−91.0) for OD and 88% (95% CI, 86.3−89.1) for OS, compared with the AUROC for H-VCDR of 81% (95% CI, 78.6−82.5) for OD and 79% (95% CI, 76.8−81.2) for OS. All differences between ML and human VCDR estimations were statistically significant. An alternative model fitted to assess the association between glaucoma status and only the baseline characteristics of age and sex explained just 2% (95% CI, 1.51−3.53) of the variance for OD and 3% (95% CI, 1.61−3.62) for OS, respectively, with an AUROC of 62% (OD, 95% CI, 59.6−64.5; OS, 95% CI, 59.8−64.7) for both.

To assess the utility of VCDR measurement as a diagnostic test for glaucoma in this cohort, accuracy metrics (sensitivity, specificity, PPV, NPV) were calculated. The results are presented in Table 1, calculated using Youden's index to determine the optimal threshold, as well as with a predetermined specificity value of 0.95 to minimize overreferrals.

Figure 4 displays the level of agreement between the ML-VCDR and human graders on a per-eye basis. For eyes without glaucoma, ML-VCDR estimates were on average lower than H-VCDR (mean VCDR 0.35 vs. 0.39 for OD and 0.34 vs. 0.39 for OS). For eyes with glaucoma, ML-VCDR estimates were on average higher than H-VCDR (mean VCDR 0.55 vs. 0.49 for OD and 0.53 vs. 0.48 for OS). These findings are supported by Bland−Altman plots in Figure S2 (available at www.aaojournal.org). Machine learning−derived VCDR estimated 87 eyes with glaucoma to have a VCDR >0.8 vs. only 2 eyes graded by humans.

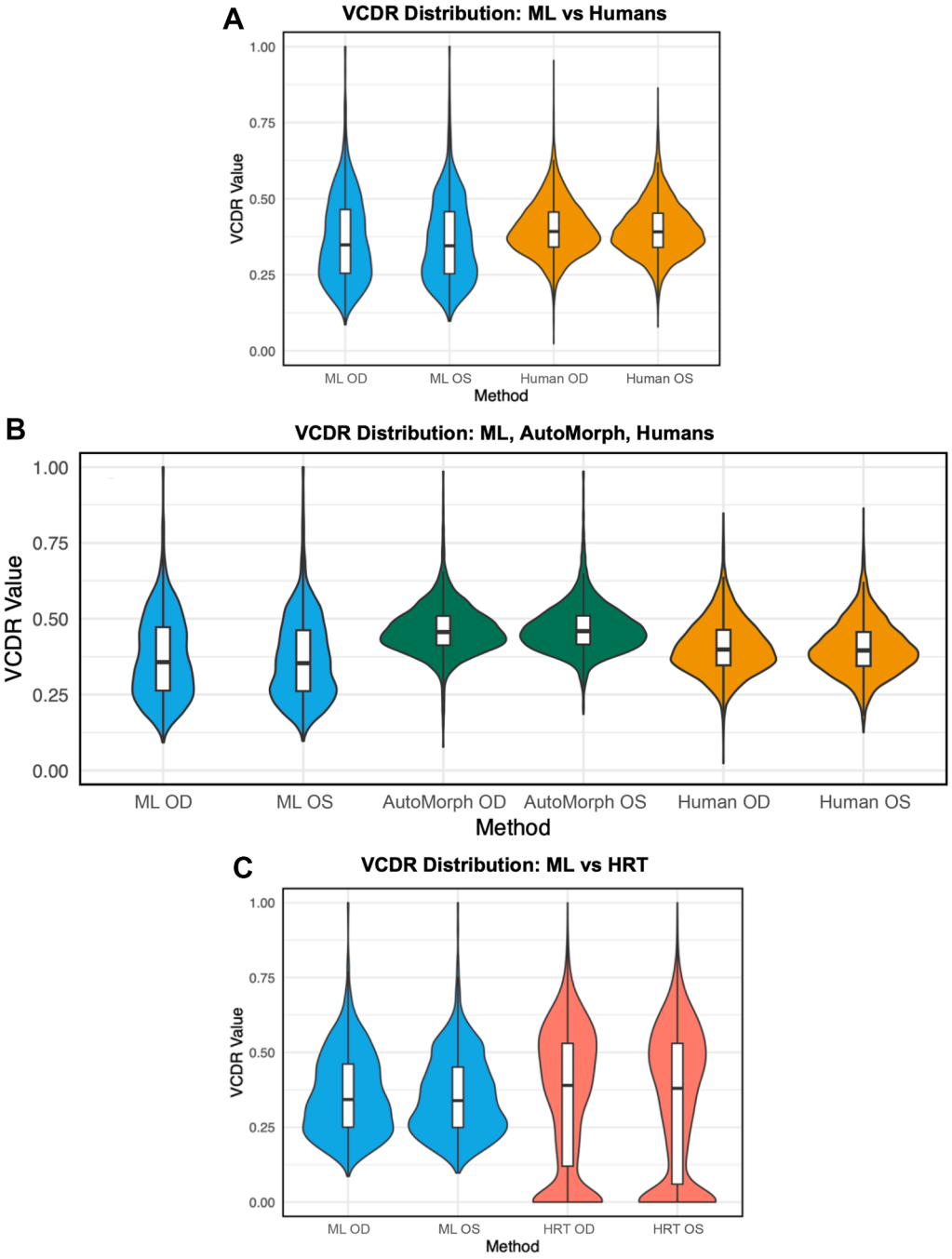


Figure 2. The distribution of vertical cup-to-disc ratios (VCDRs) for images shared between: **A,** machine-learning model (ML-VCDR) and human graders (n = 6304); **B,** machine-learning model (ML-VCDR), AutoMorph, and human graders (n = 3446); **C,** machine-learning model (ML-VCDR), and Heidelberg Retinal Tomography (HRT) (n = 4094). ML-VCDR = machine-learning-derived VCDR; OD = right eye; OS = left eye.

Subgroup Analysis Excluding Moderate and High Myopia

After exclusions based on refractive intervention and the presence of moderate or high myopia, there were 504 participants with glaucoma or suspect status in at least 1 eye in this subgroup analysis. The results were comparable with the main analysis. Machine learning–derived VCDR explained 35% (95% CI, 31.7–38.5) of variance for OD and 29% (95% CI, 26.1–32.9) for OS, whereas H-VCDR explained 21% (95% CI, 17.2–23.7) and 17% (95% CI, 14.2–20.6), respectively, in this cohort. Area under the ROC curve was 90% (OD, 95% CI, 89.0–91.7) and 88% (OS, 95% CI, 85.8–89.2) for ML-VCDR, compared with 82% (OD, 95% CI, 79.4–84.0) and 79% (OS, 95% CI, 76.9–81.9) for H-VCDR. All differences between ML and H-VCDR estimations were statistically significant.

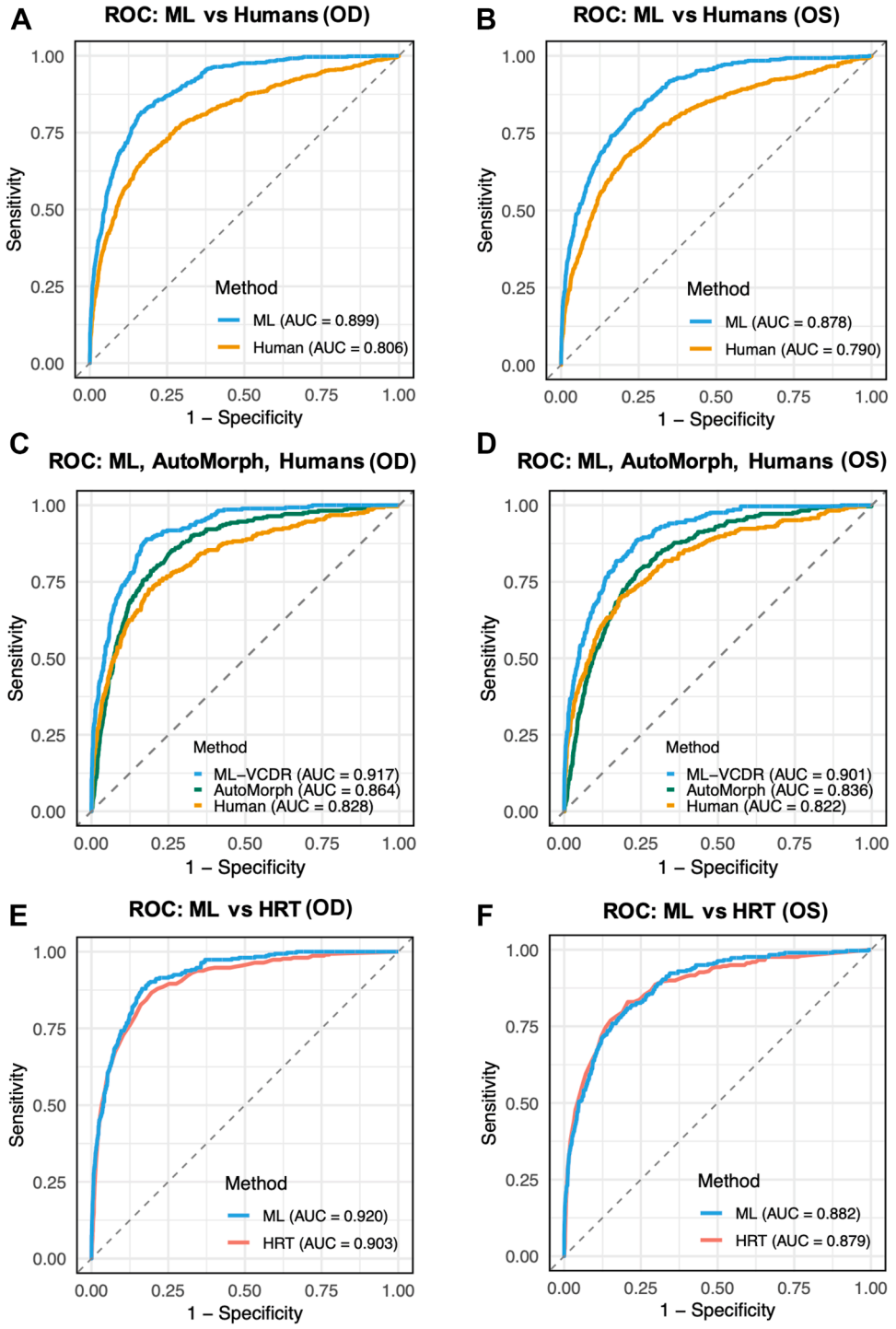


Figure 3. ROC curves for vertical cup-to-disc ratio (VCDR) as a predictor of glaucoma status, as estimated by (**A, B**) machine-learning model (ML-VCDR) and human graders (**A** = OD, **B** = OS; n = 6304); (**C, D**) machine-learning model (ML-VCDR), AutoMorph, and human graders (**C** = OD, **D** = OS; n = 3446); (**E, F**) machine-learning model (ML-VCDR) and Heidelberg Retinal Tomography (HRT) (**E** = OD, **F** = OS; n = 4094). Grey dashed lines represent the line of no discrimination (AUC = 0.50). AUC = area under the ROC curve; ML-VCDR = machine-learning-derived VCDR; OD = right eye; OS = left eye; ROC = receiver operating characteristic.

Glaucoma Risk Score

An additional metric generated by the ML model was a continuous probability score for the risk of having glaucoma (defined as being a definite or suspected case) rather than VCDR. This score had a slightly higher AUROC than VCDR alone for both OD and OS of 91% (95% CI, 90.2–92.3) and 89% (95% CI, 88.1–90.8), respectively, which was a

Table 1. Diagnostic Performance of VCDR for Predicting Glaucoma Status at the Optimal Threshold as Determined by Youden's Index and at the Threshold Corresponding to 95% Specificity

	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value	Threshold
Optimal Threshold					
ML-VCDR OD	81.6 (78.4–86.5)	83.4 (78.8–85.1)	33.3 (28.4–36.4)	97.8 (97.4–98.4)	0.48 (0.45–0.49)
H-VCDR OD	68.9 (62.7–79.7)	80.1 (70.3–85.1)	26.1 (20.4–31.3)	96.2 (95.6–97.3)	0.46 (0.43–0.47)
ML-VCDR OS	**81.3 (73.4–92.4)**	77.4 (65.4–84.4)	26.5 (20.1–33.2)	**97.6 (96.9–98.9)**	0.44 (0.39–0.48)
H-VCDR OS	67.4 (62.3–78.5)	79.3 (67.9–83.4)	24.7 (19.1–27.9)	96.0 (95.6–97.0)	0.45 (0.42–0.46)
Threshold Corresponding to 95% Specificity					
ML-VCDR OD	**50.5 (45.9–57.0)**	95.0	**50.7 (47.2–54.2)**	**95.0 (94.4–95.7)**	**0.56 (0.55–0.57)**
H-VCDR OD	39.0 (34.9–43.5)	95.0	44.2 (40.8–47.5)	93.9 (93.3–94.5)	0.53 (0.53–0.54)
ML-VCDR OS	**48.8 (43.6–52.8)**	95.0	**49.5 (45.8–52.3)**	**94.9 (94.2–95.5)**	**0.55 (0.55–0.56)**
H-VCDR OS	33.0 (29.3–37.8)	95.0	39.9 (36.2–43.6)	93.4 (92.7–94.1)	0.53 (0.53–0.54)

Point estimates for the entire data set with bootstrapped 95% confidence interval. Bold values indicate a statistically significant difference between ML and human VCDR estimates per eye.
H-VCDR = human-grader estimate of VCDR; ML-VCDR = machine-learning-derived VCDR; OD = right eye; OS = left eye; VCDR = vertical cup-to-disc ratio.

statistically significant difference. Diagnostic performance criteria, including sensitivity, specificity, PPV, and NPV, are illustrated for the referability score as a diagnostic test in Table S2 (available at www.aaojournal.org) at the optimal threshold as determined by Youden's index and at a set specificity threshold of 95%. Calibration curves for the risk scores are shown in Figure S3 (available at www.aaojournal.org) and indicate an underestimation of risk, with predicted probabilities generally lower than observed glaucoma proportions across the risk spectrum.

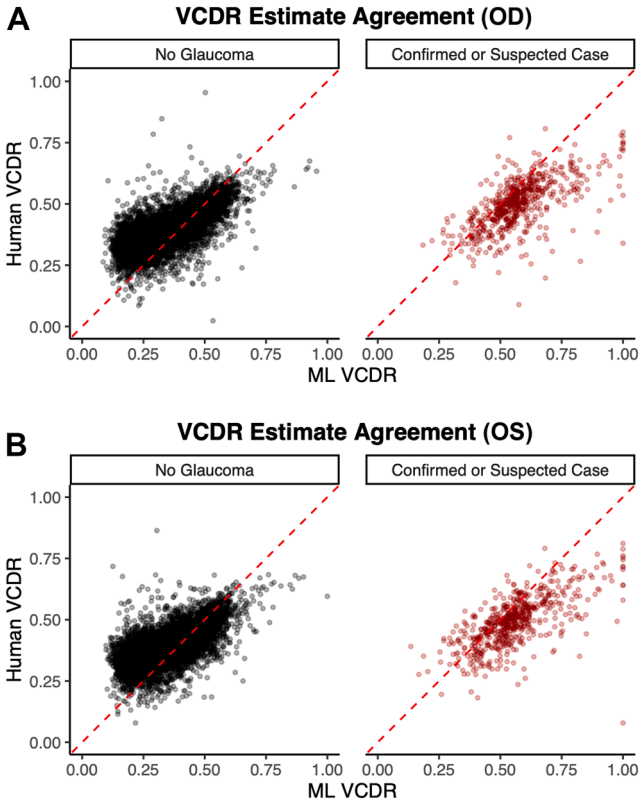


Figure 4. Intergrader VCDR estimate agreement: (**A**) right eyes, (**B**) left eyes. Red dashed line: x = y. Human = human-grader estimate of VCDR; ML = machine-learning-derived VCDR; OD = right eye; OS = left eye; VCDR = vertical cup-to-disc ratio.

ML-VCDR vs. Human Graders vs. AutoMorph

In total, 3446 participants had images that were gradable by the ML model trained for this study (ML-VCDR), AutoMorph, and human graders in both eyes. The smaller sample size reflects the fact that AutoMorph also performs automated analysis of retinal vascular morphology and requires higher-quality images to do so. The distribution of predicted VCDR results between the groups is shown in Figure 2B. A total of 343 participants had glaucoma or suspect status in at least 1 eye (280 OD and 286 OS). The ML-VCDR prediction consistently explained the most variance and had the highest AUROC, as shown in Table 2 and Figure 3C, D.

ML-VCDR vs. SLO

In total, 4094 participants had images that were gradable in both eyes by ML-VCDR and HRT. The distribution of predicted VCDR results is shown in Figure 2C. Heidelberg Retinal Tomography estimated far more VCDRs to be exactly 0 than ML-VCDR. This is a known phenomenon with HRT assessment relating to how the reference plane is set in relation to the disc height at the temporal margin.[23] Machine learning–derived VCDR also estimated these cases to have small VCDRs (mean = 0.23). The ML-VCDR prediction again explained the most variance and had slightly higher AUROC values, as shown in Table 2 and Figure 3E, F. The difference was statistically significant for OD, but not for the left.

Discussion

In this population-based setting, an ML model outperformed trained human graders at predicting specialist-confirmed glaucoma status, using VCDR as the sole diagnostic test, and explained a greater degree of the variance in glaucoma

diagnosis. The ML model trained for this study (ML-VCDR) also outperformed both an established, open-access ML model (AutoMorph) and SLO (HRT) in the prediction of glaucoma status based on VCDR. There was an additional, small improvement in predictive ability when comparing the ML-generated overall glaucoma referability score with VCDR alone. This may indicate that the model primarily derives its predictive value from VCDR, with minor contribution from other potential features such as focal notching, nerve fiber layer defects, and disc hemorrhages; however, this is speculative.

A significant strength of this study is that the ML predictions on EPIC-Norfolk images were generated without prior exposure to these images, establishing a rigorous external validation. The strong performance suggests the model's potential for broad generalizability to diverse, independent sources. By making the model openly accessible, we anticipate that other researchers will be able to evaluate the extent of this generalizability in their own cohorts. Furthermore, as the validation data set is from a population study with a disease prevalence and case mix indicative of the general population, it is a closer representation of how such a model may perform in a true screening scenario.

Whether the model's performance is sufficient for use in screening is a complex issue. National screening committees do not specify minimum performance levels for screening tests.[24] The suitability of a test for screening depends on a range of factors, including but not limited to disease prevalence, the cost of a test, harms associated with subsequent diagnostic testing, and the consequences of unidentified cases. A review of mammography test performance for breast cancer screening found wide-ranging estimates of 27% to 97% for sensitivity and 46% to 99% for specificity.[25] Smear test performance for cervical cancer is similarly variable, with reported estimates between 44% and 82% sensitivity and 88% and 99% specificity.[26] The results for ML-VCDR fall within these parameters, which, at a minimum, suggests plausibility for use in screening.

Table 2. Variance Explained and AUROC Values for VCDR as a Predictor of Glaucoma Status, as Determined by ML-VCDR, AutoMorph, Human Graders, and HRT

	Variance Explained	AUROC
ML-VCDR vs. AutoMorph vs. Human Graders (n = 3446)		
ML-VCDR OD	39.2 (35.0–43.0)	91.7 (90.2–93.0)
ML-VCDR OS	35.2 (31.3–39.2)	90.1 (88.3–91.7)
AM-VCDR OD	21.8 (18.4–25.7)	86.4 (84.2–88.4)
AM-VCDR OS	16.5 (13.4–19.9)	83.6 (81.4–85.8)
H-VCDR OD	22.6 (19.0–26.8)	82.8 (80.1–85.4)
H-VCDR OS	22.4 (18.6–26.5)	82.3 (79.5–84.8)
ML-VCDR vs. HRT (n = 4094)		
ML-VCDR OD	39.6 (35.7–43.6)	92.0 (90.6–93.4)
ML-VCDR OS	31.9 (28.1–35.9)	88.2 (86.3–90.0)
HRT-VCDR OD	33.4 (29.1–38.1)	90.3 (88.5–92.0)
HRT-VCDR OS	28.7 (24.3–33.2)	87.9 (85.6–90.0)

Point estimates for the entire data set with bootstrapped 95% confidence interval.
AM-VCDR = AutoMorph estimate of VCDR; AUROC = area under the receiver operating characteristic curve; HRT-VCDR = Heidelberg Retinal Tomography estimate of VCDR; H-VCDR = human-grader estimate of VCDR; ML-VCDR = machine-learning model-derived VCDR trained for this study; OD = right eye; OS = left eye; VCDR = vertical cup-to-disc ratio.

The encouraging performance of the model in this external validation likely relates, in part, to the strength of its training data. The foundation of ML-VCDR is a model trained on a large number of ophthalmologist-labeled images.[19] The expertise captured by these labels, in combination with the additional large-scale training on UKB images, provided a high level of both quality and quantity for the model's training input, improving both the performance and generalizability of the model.

A further strength of the study is the method of glaucoma status ascertainment in EPIC-Norfolk. Definitive diagnosis was made through direct participant examination by a consultant ophthalmologist specializing in glaucoma.[18] This specialist had access to a full range of clinical tests in addition to this and could repeat investigations if required. Additionally, a second consultant ophthalmologist specializing in glaucoma reviewed all examination findings and history in a subgroup of participants to minimize falsely positive or negative diagnoses.[18] This comprehensive clinical assessment incorporated the range of investigations recommended for glaucoma diagnosis.[27,28] Other studies assessing AI performance for glaucoma diagnosis often rely on optic disc assessment alone to establish the ground truth for glaucoma status.[29,30] As such, these studies are in fact assessing the detection of optic nerve head appearances that may be consistent with glaucoma, not their overall glaucoma status. This gives our study greater external validity because of more clinically accepted glaucoma ascertainment. It also, at least in part, explains why the model performance is poorer than in those studies that only include disc assessment; the ML-VCDR model uses only 1 investigation among the many that are used to form the definitive diagnosis in EPIC-Norfolk. The EPIC-Norfolk Eye Study is a valuable resource for such analyses as it combines population-based data with reliable glaucoma status ascertainment. In contrast, some large-scale population studies rely on self-reported data or proxy measures, including questionnaires or medication use, to ascertain disease status.[31,32]

A recent study by Chaurasia et al [33] also trained an ML model for cup-disc ratio estimation on UKB images and validated it for glaucoma prediction on an external data set. Our study benefits from glaucoma status being confirmed by direct participant examination as opposed to human review of fundus photos only. We also compare the performance of our model with human estimates of VCDR in the external test set. By doing so, our study shows the extent to which ML-VCDR can improve on human-estimated VCDR in a screening scenario with reliable case ascertainment.

It is noteworthy that ML-VCDR estimated far more eyes to have a VCDR >0.8 than the human graders. This could indicate that human graders systematically underestimated the VCDRs. A potential explanation is that H-VCDR

measurement from color fundus photography is a fundamentally difficult task, susceptible to subjectivity and variability, which may lead to caution.[34,35] Direct, on-screen measurements require precise, consistent decision-making regarding boundaries, and ML models may well be better at this than humans and with less fatigue.

Although the ML estimate of VCDR is more *predictive* of glaucoma than the human graders' estimates, it is not necessarily a *truer* representation. All reasonable methods for attempting to define a ground truth for VCDR require some form of subjectivity when defining boundaries. Machine learning–derived VCDR does so in a way that better captures the risk of glaucoma of that disc than the human graders do.

This study has some limitations. All color fundus images were 2 dimensional, and it is possible that the performance of both ML-VCDR and humans may have been better with stereoscopic assessment.[36] Furthermore, EPIC-Norfolk has an ancestrally homogeneous participant cohort, with almost all being of European ancestry.[17] Therefore, we do not know how the model's performance might vary in individuals from other ancestral groups, although our prior study did not detect population-specific effects on VCDR prediction quality.[19]

These observations support the idea that ML models could be incorporated into glaucoma assessment to improve diagnostic accuracy. Machine learning–derived VCDR performs better than human graders at identifying glaucoma based on VCDR and does so in a community-based population, with glaucoma status based on direct participant examination, mimicking a screening scenario. Because ML-based assessment of fundus photographs has the potential to be significantly faster and more cost-effective than human assessment, it may, with further development and evaluation, represent a viable approach for population-level screening.

Acknowledgments

This research has been conducted using the UK Biobank Resource under Application Number 36741.

Footnotes and Disclosures

Originally received: February 20, 2026.
Final revision: July 31, 2026.
Accepted: July 31, 2026.
Available online: August 10, 2026. Manuscript no. OPHTHA-D-26-00470.

[1] NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, UK.

[2] Google Research, Mountain View, California.

[3] Department of Sense Organs, Sapienza University of Rome, Rome, Italy.

[4] Norfolk & Norwich University Hospital, Norfolk, UK.

[5] Singapore National Eye Centre & Singapore Eye Research Institute, Singapore.

[6] Centre for Population Health Research & Implementation (CPHRI), SingHealth, Singapore.

[7] Google Research, Cambridge, Massachusetts.

*T.R.P.T., J.K., and R.N.L. contributed equally as first authors.

†C.Y.M. and A.P.K. contributed equally as senior authors.

Disclosure(s):

The Article Publishing Charge (APC) for this article was paid by University College London.

All authors have completed and submitted the ICMJE disclosures form.

The authors made the following disclosures:

J.K.: Employer – Google LLC.

F.H.: Employer – Google LLC.

A.C.: Employer – Google LLC.

C.Y.M.: Employer – Google LLC.

L.A.M.: Writing fees – Santen.

D.J.F.: Consulting fees/research support – Abbvie, Allergan, Apellis, Boehringer Ingelheim, Galimedix, Google Health, and Roche.

P.A.K.: Cofounder – Cascader Ltd; Consultant – Skleo Health, insitro, Retina Consultants of America, Roche, Boehringer Ingelheim, Bitfount; Equity owner – Big Picture Medical; Speaker fees – Zeiss, Thea, Apellis, Roche; Grant funding – Roche; Travel support – Bayer and Roche. Attended advisory board –Topcon, Bayer, Boehringer Ingelheim, Roche.

P.J.F.: Consulting fees – Alphasights, GLG, Google Health, Guidepoint, PwC, Santen; Lecture fees – Viatris; Data monitoring board fees – Abbvie.

A.P.K.: Consulting fees – Abbvie, Aerie, Allergan, Glaucore, Google Health, Heidelberg, Novartis, Regeneron, Reichert, Santen, Seonix Bio, Thea, Topcon.

Supported by a grant from the National Institute for Health and Social Care Research (NIHR) for a Biomedical Research Centre (BRC) at Moorfields Eye Hospital NHS Foundation Trust and UCL (University College London) Institute of Ophthalmology. The EPIC-Norfolk Study (https://doi.org/10.22025/2019.10.105.00004) has received funding from the Medical Research Council (MR/N003284/1 MC-UU_12015/1 and MC_UU_00006/1) and Cancer Research UK (C864/A14136). The clinic for the third health examination was funded by Research into Ageing (262). T.R.P.T. is supported by the National Institute for Health and Care Research (NIHR) Moorfields Biomedical Research Centre (BRC4-02-RB402-117) and the Lister Institute. L.A.M. is supported by a Marco Centofanti Research Award 2024-2025 OMIKRON Italia/Italian Association for the Study of Glaucoma (A.I.S.G.). K.V.S. is supported by Fight for Sight (UK, 1956A) and The Desmond Foundation, and a National Institute for Health and Care Research (NIHR) Moorfields Biomedical Research Centre Career Enhancement Award. D.J.F. is supported by NIHR and the Wellcome Trust. P.A.K. is supported by a UK Research & Innovation Future Leaders Fellowship (MR/T019050/1), Moorfields Eye Charity with The Rubin Foundation Charitable Trust (GR001753), and an Alcon Research Institute Senior Investigator Award. P.J.F. is supported by an NIHR grant to Moorfields Eye Hospital NHS Foundation Trust and University College London (UCL) Institute of Ophthalmology for a Biomedical Research Centre (BRC4), Fight for Sight and The Desmond Foundation. A.P.K. is supported by a UK Research and Innovation (UKRI) Future Leaders Fellowship (MR/T040912/1), Moorfields Eye Charity Career Development Fellowship, and Lister Institute of Preventative Medicine Fellowship. The funding organizations had no role in the design or conduct of this research.

For the purposes of open access, the author has applied a Creative Commons Attribution (CC BY) license to any author-accepted manuscript version arising.

Data Availability: Data may be available by reasonable request to the corresponding author, subject to the approval of contributing studies.

Presented by Professor Anthony Khawaja at the American Academy of Ophthalmology Annual Meeting 2025 (PA004).

Anthony P. Khawaja, PhD, FRCOphth, an editorial board member of this journal, was recused from the peer-review process of this article and had no access to information regarding its peer-review.

HUMAN SUBJECTS: Human subjects were used in this study. The study was approved by the Norwich Local Research Ethics Committee (05/Q0101/191) and the East Norfolk & Waveney NHS Research Governance Committee (2005EC07L). All research adhered to the tenets of the Declaration of Helsinki. All participants provided informed consent.

No animal subjects were used in this study.

Author Contributions:

Conception and design: Taylor, Khasentino, McLean, Khawaja.

Data collection: Taylor, Khasentino, Luben, Chan, Broadway, Hormozdiari, Carroll, Foster, McLean, Khawaja.

Analysis and interpretation: Taylor, Khasentino, Luben, Meliante, Stuart, Fu, Hu, Verburg, Biradar, Keane, Foster, McLean, Khawaja.

Obtained funding: Foster, Khawaja.

Overall responsibility: Taylor, Khasentino, Luben, Meliante, Stuart, Fu, Chan, Broadway, Hormozdiari, Hu, Verburg, Biradar, Keane, Foster, McLean, Khawaja.

Abbreviations and Acronyms:
AUROC = area under the receiver operating characteristic curve; **CI** = confidence interval; **EPIC** = European Prospective Investigation of Cancer; **HRT** = Heidelberg Retinal Tomography; **H-VCDR** = human-derived vertical cup-disc ratio; **IQR** = interquartile range; **ML** = machine learning; **ML-VCDR** = machine learning–derived vertical cup-disc ratio; **NPV** = negative predictive value; **OD** = right eye; **OS** = left eye; **PPV** = positive predictive value; **ROC** = receiver operating characteristic; **SLO** = scanning laser ophthalmoscopy; **UKB** = UK Biobank; **VCDR** = vertical cup-disc ratio.

Keywords:
Glaucoma, Machine learning, Screening, VCDR.

Correspondence:
Thomas R.P. Taylor, BMBCh, Institute of Ophthalmology, University College London, 11-43 Bath Street, London EC1V 9EL, UK. E-mail: thomas.taylor.25@ucl.ac.uk.

References

1. GBD 2019 Blindness and Vision Impairment Collaborators. Vision Loss Expert Group of the Global Burden of Disease Study. Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the Right to Sight: an analysis for the Global Burden of Disease Study. *Lancet Global Health*. 2021;2:E144–E160. Accessed December 9, 2025 https://www.thelancet.com/journals/langlo/article/PIIS2214-109X(20)30489-7/fulltext.
2. Jones PR, Philippin H, Makupa WU, et al. Severity of visual field loss at first presentation to glaucoma clinics in England and Tanzania. *Ophthalmic Epidemiol*. 2020;27:10–18. https://doi.org/10.1080/09286586.2019.1661499
3. Wong TL, Ang JL, Deol S, et al. The relationship between multiple deprivation and severity of glaucoma at diagnosis. *Eye*. 2023;37:16. https://doi.org/10.1038/s41433-023-02508-w
4. Stuart KV, de Vries VA, Schuster AK, et al. Prevalence of glaucoma in Europe and projections to 2050: findings from the European Eye Epidemiology Consortium. *Ophthalmology*. 2025;132:1114–1124. https://doi.org/10.1016/j.ophtha.2025.06.002
5. Fraser S, Bunce C, Wormald R, Brunner E. Deprivation and late presentation of glaucoma: case-control study. *BMJ*. 2001;322:639–643. https://doi.org/10.1136/bmj.322.7287.639
6. UK National Screening Committee. Glaucoma-UK National Screening Committee (UK NSC) - GOV.UK. https://view-health-screening-recommendations.service.gov.uk/glaucoma/. Accessed October 19, 2023.
7. US Preventive Services Task Force, Mangione CM, Barry MJ, et al. Screening for primary open-angle glaucoma: US Preventive Services Task Force recommendation statement. *JAMA*. 2022;327:1992. https://doi.org/10.1001/jama.2022.7013
8. Burr J, Hernández R, Ramsay C, et al. Is it worthwhile to conduct a randomized controlled trial of glaucoma screening in the United Kingdom? *J Health Serv Res Policy*. 2014;19:42–51. https://doi.org/10.1177/1355819613499748
9. Azuara-Blanco A, Burr J, Thomas R, et al. The accuracy of accredited glaucoma optometrists in the diagnosis and treatment recommendation for glaucoma. *Br J Ophthalmol*. 2007;91:1639–1643. https://doi.org/10.1136/bjo.2007.119628
10. Chaurasia AK, Greatbatch CJ, Hewitt AW. Diagnostic accuracy of artificial intelligence in glaucoma screening and clinical practice. *J Glaucoma*. 2022;31:285–299. https://doi.org/10.1097/IJG.0000000000002015
11. Wu JH, Nishida T, Weinreb RN, Lin JW. Performances of machine learning in detecting glaucoma using fundus and retinal optical coherence tomography images: a meta-analysis. *Am J Ophthalmol*. 2022;237:1–12. https://doi.org/10.1016/j.ajo.2021.12.008
12. Gao XR, Wu F, Yuhas PT, Rasel RK, Chiariglione M. Automated vertical cup-to-disc ratio determination from fundus images for glaucoma detection. *Sci Rep*. 2024;14:4494. https://doi.org/10.1038/s41598-024-55056-y
13. Fu H, Cheng J, Xu Y, et al. Disc-aware ensemble network for glaucoma screening from fundus image. *IEEE Trans Med Imaging*. 2018;37:2493–2501. https://doi.org/10.1109/TMI.2018.2837012
14. Jiang Y, Duan L, Cheng J, et al. JointRCNN: a region-based convolutional neural network for optic disc and cup segmentation. *IEEE Trans Biomed Eng*. 2020;67:335–343. https://doi.org/10.1109/TBME.2019.2913211
15. Riboli E, Kaaks R. The EPIC Project: rationale and study design. European Prospective Investigation into Cancer and Nutrition. *Int J Epidemiol*. 1997;26(suppl 1):S6–S14. https://doi.org/10.1093/ije/26.suppl_1.s6
16. Day N, Oakes S, Luben R, et al. EPIC-Norfolk: study design and characteristics of the cohort. European Prospective Investigation of Cancer. *Br J Cancer*. 1999;80(suppl 1):95–103.
17. Khawaja AP, Chan MPY, Hayat S, et al. The EPIC-Norfolk Eye Study: rationale, methods and a cross-sectional analysis of visual impairment in a population-based cohort. *BMJ Open*. 2013;3:e002684. https://doi.org/10.1136/bmjopen-2013-002684

18. Chan MPY, Broadway DC, Khawaja AP, et al. Glaucoma and intraocular pressure in EPIC-Norfolk Eye Study: cross sectional study. *BMJ*. 2017;358:j3889. https://doi.org/10.1136/bmj.j3889
19. Alipanahi B, Hormozdiari F, Behsaz B, et al. Large-scale machine-learning-based phenotyping significantly improves genomic discovery for optic nerve head morphology. *Am J Hum Genet*. 2021;108:1217−1230. https://doi.org/10.1016/j.ajhg.2021.05.004
20. Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision. arXiv. *In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016:2818−2826. https://doi.org/10.48550/arXiv.1512.00567
21. Deng J, Dong W, Socher R, et al. In: *ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition*. Miami, FL, USA: IEEE; 2009:248−255. https://doi.org/10.1109/CVPR.2009.5206848
22. Zhou Y, Wagner SK, Chia MA, et al. AutoMorph: automated retinal vascular morphology quantification via a deep learning pipeline. *Transl Vis Sci Technol*. 2022;11:12. https://doi.org/10.1167/tvst.11.7.12
23. Khawaja AP, Chan MPY, Broadway DC, et al. Laser scanning tomography in the EPIC-Norfolk Eye Study: principal components and associations. *Invest Ophthalmol Vis Sci*. 2013;54:6638−6645. https://doi.org/10.1167/iovs.13-12490
24. GOV.UK. Criteria for a population screening programme - GOV.UK. https://www.gov.uk/government/publications/evidence-review-criteria-national-screening-programmes/criteria-for-appraising-the-viability-effectiveness-and-appropriateness-of-a-screening-programme. Accessed March 26, 2026.
25. Fitzjohn J, Zhou C, Chase JG. Critical assessment of mammography accuracy. *IFAC-Pap*. 2023;56:5620−5625. https://doi.org/10.1016/j.ifacol.2023.10.472
26. Chen C, Yang Z, Li Z, Li L. Accuracy of several cervical screening strategies for early detection of cervical cancer: a meta-analysis. *Int J Gynecol Cancer*. 2012;22:908−921. https://doi.org/10.1097/IGC.0b013e318256e5e4
27. Primary Open-Angle Glaucoma PPP 2020. American Academy of Ophthalmology. November 13, 2020. https://www.aao.org/education/preferred-practice-pattern/primary-open-angle-glaucoma-ppp. Accessed January 2, 2026.
28. Recommendations | Glaucoma: diagnosis and management | Guidance | NICE. November 1, 2017. https://www.nice.org.uk/guidance/ng81/chapter/Recommendations#diagnosis. Accessed January 2, 2026.
29. Christopher M, Belghith A, Bowd C, et al. Performance of deep learning architectures and transfer learning for detecting glaucomatous optic neuropathy in fundus photographs. *Sci Rep*. 2018;8:16685. https://doi.org/10.1038/s41598-018-35044-9
30. Li Z, He Y, Keel S, et al. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. *Ophthalmology*. 2018;125:1199−1206. https://doi.org/10.1016/j.ophtha.2018.01.023
31. Grant A, Roy-Gagnon MH, Bastasic J, et al. Exploring ethnic and racial differences in intraocular pressure and glaucoma: the Canadian Longitudinal Study on aging. *Heliyon*. 2024;10:e28611. https://doi.org/10.1016/j.heliyon.2024.e28611
32. Neustaeter A, Vehof J, Snieder H, Jansonius NM. Glaucoma in large-scale population-based epidemiology: a questionnaire-based proxy. *Eye*. 2021;35:508−516. https://doi.org/10.1038/s41433-020-0882-4
33. Chaurasia AK, Greatbatch CJ, Han X, et al. Highly accurate and precise automated cup-to-disc ratio quantification for glaucoma screening. *Ophthalmol Sci*. 2024;4:100540. https://doi.org/10.1016/j.xops.2024.100540
34. Kahn HA, Leibowitz H, Ganley JP, et al. Standardizing diagnostic procedures. *Am J Ophthalmol*. 1975;79:768−775. https://doi.org/10.1016/0002-9394(75)90735-7
35. Varma R, Steinmann WC, Scott IU. Expert agreement in evaluating the optic disc for glaucoma. *Ophthalmology*. 1992;99:215−221. https://doi.org/10.1016/s0161-6420(92)31990-6
36. Lehmann MV, Mardin CY, Martus P, Bergua A. 3D vs 2D qualitative and semiquantitative evaluation of the glaucomatous optic disc atrophy using computer-assisted stereophotography. *Eye*. 2008;22:628−635. https://doi.org/10.1038/sj.eye.6702690