Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Evaluating medical AI models beyond predictive performance: a diabetic retinopathy screening case study

Hyunmin Kim
University College London
Rahul A. Jonas
University of Cologne
Takahiro Ninomiya
Tohoku University
Rimkyung Hong
Hanyang University Guri Hospital
Vasilena Sitnilska
University of Cologne
Antonia Howaldt
University College London
Ariel Yuhan ONG
University College London
Michael Grün
St. Franziskus Hospital
Broder Poschkamp
Greifswald University Hospital
Yukun Zhou
University College London
Lie Ju
University College London
Yiqun Lin
University College London
Eden Ruffell
University College London
Dominic Williamson
University College London
Josef Huemer
University College London
Koichi Ito
Tohoku University
Nderitu Paul
University College London
Michael G. Morley
Massachusetts Eye and Ear, Harvard Medical School
Mohammad Eslami
Massachusetts Eye and Ear, Harvard Medical School
Elze Tobias
Massachusetts Eye and Ear, Harvard Medical School
Eric F. Thee
Erasmus MC
Junseo Choi
Hanyang University Guri Hospital
Danbi Lee
Hanyang University Guri Hospital
Yaxing Wang
Tsinghua University
Songhomitra Panda-Jonas
Rothschild Foundation Hospital
Jost B. Jonas

Rothschild Foundation Hospital
Mukharram M. Bikbov
Ufa Eye Research Institute
Gyulli M. Kazakbaeva
Ufa Eye Research Institute
Vinay Nangia
Suraj Eye Institute
Siegfried K. Wagner
University College London
Sunjin Hwang
Hanyang University Guri Hospital
Pearse A. Keane
p.keane@ucl.ac.uk

University College London

Article

Keywords:

Posted Date: June 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9848475/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: Competing interest reported. P.A.K. is a co-founder of Cascader Ltd. and has acted as a consultant for Skleo Health, insitro, Retina Consultants of America, Roche, Boehringer-Ingelheim, and Bitfount, and is an equity owner in Big Picture Medical. He has received speaker fees from Zeiss, Théa, Apellis, and Roche, travel support from Bayer and Roche, and grant funding from Roche. He has attended advisory boards for Topcon, Bayer, Boehringer-Ingelheim, and Roche.

Abstract

Predictive metrics are widely used to evaluate artificial intelligence (AI)-based screening tools, yet their clinical reliability requires a more comprehensive assessment across multiple aspects, including robustness and calibration. Using diabetic retinopathy (DR) screening as an exemplar case study, we simulated a model development and update scenario to reflect realistic deployment requirements, by incrementally developing models using a development cohort of 63,301 fundus images. We evaluated the models on six external validation datasets across four reliability dimensions: probability calibration, uncertainty estimation, robustness to acquisition noise and device-induced variability, and consistency across model updates. Although predictive performance remained high across all models and all six external datasets, with a median AUROC of 0.974 (0.904–0.993), the performance of the four clinical reliability measures is limited. Model calibration was poor (Expected Calibration Error; 0.031–0.133), and estimated uncertainty did not align with clinicians' assessments (Pearson's *r*; 0.156–0.548). Although robustness was generally acceptable (Cohen's kappa; 0.431–0.852), it was reduced in borderline cases (Cohen's kappa; −0.027–0.331). Notably, model updates also introduced decision instability in borderline cases (Cohen's kappa; 0.060–0.910) and substantial fluctuations in the Youden-optimal threshold across external datasets (0.107–0.799). These findings highlight that a comprehensive evaluation across multiple complementary dimensions is required to ensure the reliability of AI-based screening tools. To support further research, we publicly released the HMADR_Clinic dataset, comprising 766 fundus photographs and their DR grades established through a hierarchical grading protocol involving five independent ophthalmologists and senior adjudication.

Introduction

Artificial intelligence (AI)-based screening tools are increasingly developed and commercialised with the promise of improving efficiency, expanding access to care, and reducing clinicians' workload.[1–3] In many areas of medical AI, such tools continue to be evaluated primarily using predictive performance metrics such as the area under the receiver operating characteristic curve (AUROC), sensitivity, and specificity.[4–6]

However, superior predictive performance alone cannot ensure reliable clinical deployment of screening tools. For example, screening tools should be robust across populations and device-related variation. Even when images are acquired from the same patient during a single visit, image quality may vary due to artefacts such as glare, light leakage, blur, or attenuation.[7,8] Device-induced variability, including differences across cameras and scanners, has received increasing attention, yet its impact is typically assessed at the population level rather than for individual images.[9] Furthermore, consistency across successive model updates remains largely overlooked; most evaluations are still conducted as one-off, static assessments, even though model updates are inevitable during post-deployment maintenance.[6,10–12]

Moreover, current AI products integrated into clinical workflows predominantly provide categorical outputs, such as positive (referral) or negative (non-referral) decisions.[13] However, the role of predicted probabilities is expected to expand beyond such categorical outputs, enabling clinicians to assess model confidence at the level of individual decisions. For these probabilities to be clinically meaningful, two conditions should be satisfied. First, they should reflect clinical ambiguity by assigning greater uncertainty to cases that clinicians themselves find difficult to classify.[14,15] Second, they should be well calibrated. That is, the probabilities the model outputs should accurately reflect true event rates in the target population, such that a predicted probability of 0.8 genuinely corresponds to an 80% likelihood of referable disease. Poor calibration may distort clinical judgement, either by inflating confidence in uncertain predictions or by undermining trust in reliable ones.[16–18]

Diabetic retinopathy (DR) screening provides a particularly relevant context for examining these issues, as it is among the few medical AI applications to have achieved regulatory clearance, commercial deployment, and adoption within national health systems across multiple countries.[19–21] Consequently, DR screening has progressed beyond proof-of-concept model development to a stage where post-market surveillance and continual model maintenance are increasingly important.[22] Reflecting this shift, regulatory bodies are placing greater emphasis on lifecycle considerations for AI-enabled medical devices.[11,12] These characteristics make DR screening a particularly informative test case for assessing whether strong predictive performance alone is sufficient for reliable deployment.

In this study, we use DR screening as a representative use case to systematically evaluate four underexamined but deployment-critical dimensions of screening AI using six external validation datasets. The dimensions of screening AI we focused are: probability calibration, alignment between model-derived and clinician-derived uncertainty, robustness to acquisition noise and cross-device variability, and consistency across successive model updates. We observed several key findings: 1) a high predictive performance does not guarantee reliable behaviour across deployment-relevant dimensions; 2) model outputs may fluctuate at individual image level in response to image noise, device variation, and model updates; and 3) although a modest association exists between model-derived uncertainty and clinician-perceived uncertainty, the two perspectives may not always align. Beyond these findings, we publicly release a new fundus image dataset of 766 images from Hanyang University Guri Hospital (HMADR_Clinic [Hanyang Multi Annotator Diabetic Retinopathy]), independently annotated by five ophthalmologists and adjudicated by a senior retinal specialist, to support uncertainty estimation benchmarking in DR screening. An overview of the study design is presented in Fig. 1.

Results

Study Population

A total of 63,301 images from seven public datasets were used in the development cohort. Four datasets (EyePACS, BRSET, mBRSET, DeepDRiD) were partitioned into training, tuning, and validation sets at the patient level, while the remaining three (APTOS 2019, MESSIDOR2, DDR) were used entirely for

training, as patient-level identifiers were unavailable. The training set comprised 43,687 images (9,113 referable-DR), the tuning set included 9,810 images (1,465 referable-DR), and the internal validation set consisted of 9,804 images (1,440 referable-DR).

External validation was conducted on 5,213 images (796 referable-DR cases), drawn from six independent datasets. These were categorised as clinic-based datasets (IDRiD [image-level referral rate, 61.2%] and HMADR_Clinic [32.0%]), population-based epidemiological eye studies (Ural Eye Medical Study [UEMS; 6.0%], Central India Eye Medical Study [CIEMS; 3.7%], and Beijing Eye Study [BES; 8.0%]), and a population-level screening dataset (HMADR_Screening; 5.0%). All external validation datasets were filtered to include only patients with diabetes mellitus, in line with the intended clinical application of DR screening. Analyses were performed across three groupings: (1) all six datasets combined (all-datasets group); (2) eye clinic-based datasets (clinic-based group); and (3) population-based datasets (population-based group, comprising epidemiological studies and the population-level screening dataset).

Paired images from the AI-READI dataset were used to assess robustness to acquisition noise, which was defined as variability captured through intra-device pairs, and inter-device variability. This dataset comprised 3,588 intra-device pairs (Triton–Triton) for noise robustness assessment, as well as 3,887 Triton–Maestro2 pairs, 1,873 Triton–Eidon pairs, and 1,874 Maestro2–Eidon pairs for inter-device comparisons. The characteristics of the development and external validation datasets are summarised in Tables 1 and 2, respectively.

Table 1
Summary of development datasets

Dataset	Country	Year	Setting	Device	Fixation	Number of patients	Number of Images	Split	Images after QC		Total	R r
									Non-referable	Referable		
EyePACS	United States	2015	Community-based	Multiple devices (not specified)	Macula-centred or optic-disc-centred	17,563	35,126	Training	14,418	3,118	17,536	1
								Tuning	4,774	1,058	5,832	1
								Validation	4,810	1,042	5,852	1
								Total	24,002	5,218	29,220	1
BRSET	Brazil	2023	Clinic-based	NF505(Nikon, Tokyo, Japan), CR-2 (Canon, Tokyo, Japan)	Macula-centred or optic-disc-centred	8,524	16,266	Training	8,076	510	8,586	5
								Tuning	2,682	174	2,856	6
								Validation	2,681	157	2,838	5
								Total	13,439	841	14,280	5
mBRSET	Brazil	2024	Community-based	Phelcom Eyer (Phelcom Technologies, Sao Carlos, Brazil)	Macula-centred or optic-disc-centred	1,291	5,164	Training	2,395	525	2,920	1
								Tuning	811	161	972	1
								Validation	792	175	967	1
								Total	3,998	861	4,859	1
APTOS 2019	India	2019	Community-based	Not specified	Macula-centred or optic-disc-centred	Not available	3,662	Total	1,876	1,063	2,939	3
MESSIDOR2	France	2013	Clinic-based	TRC-NW6 (Topcon, Tokyo, Japan)	Macula-centred	Not available	1,748	Training (Total)	1,273	448	1,721	2
DeepDRiD	China	2020	Screening-based	Conventional fundus camera; exact model not specified	Macula-centred and optic-disc-centred	400	1,600	Training	237	217	454	4
								Tuning	78	72	150	4
								Validation	81	66	147	4
								Total	396	355	751	4
DDR	China	2019	Clinic-based	Multiple devices (42 models from Topcon, Nikon, and Canon)	Optic disc–fovea midpoint-centred	Not available	13,673	Training (Total)	6,299	3,232	9,531	3

Referable diabetic retinopathy was defined as moderate non-proliferative diabetic retinopathy or worse (ICDR grade ≥ 2). Seven datasets were used for development. Four datasets (EyePACS, BRSET, mBRSET, and DeepDRiD) were partitioned at the patient level into training, tuning, and internal validation sets 6:2:2 ratio. The remaining three datasets (APTOS 2019, MESSIDOR2, and DDR) were used entirely for training, as patient-level identifiers were unavailable. In classified as "bad" quality by AutoMorph, an automated retinal image quality assessment tool, were excluded. The "Non-referable" and "Referable" columns indicate the number of images in each class after quality control. Referral rate (%) was computed as the proportion of referable images among all retained i within each split. "Number of patients" and "Number of images" reflect the original dataset size before quality filtering; "Not available" indicates that patient-l identifiers were not provided by the dataset.

Abbreviations: QC, quality control; ICDR, International Clinical Diabetic Retinopathy.

Table 2
Summary of external validation datasets

Dataset	Country	Setting	Device	Fixation	Number of patients	Number of Images	After DM filter	After QC	Age (mean ± SD)	Sex (Male, %)	After QC
											Non-referabl
IDRiD	India (Eye Clinic, Sushrusha Hospital)	Clinic-based	VX-10 (Kowa, Nagoya, Japan)	Macula-centred	Not available	516	516	490	Not available	Not available	190
HMADR_Clinic	South Korea (tertiary hospital)	Clinic-based	DRI OCT Triton (Topcon, Tokyo, Japan)	Macula-centred	551	1,816	904 (213)	766 (190)	59.9 ± 12.7	288, 37.6	521
HMADR_Screening	South Korea (screening centre)	Screening - based	TRC-NW8 (Topcon, Tokyo, Japan)	Macula-centred	8,371	16,376	1,173 (782)	1,110 (736)	55.0 ± 10.8	649, 58.5	1,055
UEMS	Russia (population-based epidemiological study)	Population-based	Visucam 500 (Carl Zeiss Meditec, Jena, Germany)	Macula-centred and optic-disc-centred	4,672	15,963	1,236 (372)	620 (282)	62.2 ± 8.4	162, 26.1	583
CIEMS	India (population-based epidemiological study)	Population-based	FF450 (Carl Zeiss Meditec, Jena, Germany)	Macula-centred and optic-disc-centred	3,426	12,709	546 (163)	461 (146)	52.7 ± 11.8	218, 47.3	444
BES	China (population-based epidemiological study)	Population-based	CR6-45NM (Canon, Tokyo, Japan)	Macula-centred and optic-disc-centred	3,452	13,429	1,971 (507)	1,766 (494)	62.1 ± 9.8	776, 43.9	1,624

Referable diabetic retinopathy was defined as moderate non-proliferative diabetic retinopathy or worse (ICDR grade ≥ 2). Referral rate (%) was calculated as images divided by the total number of images after quality control (QC). The "Number of patients" and "Number of images" columns refer to the original data DM filter" column indicates the subset retained after filtering for diabetes mellitus based on an HbA1c level of 6.5% or higher. Image quality was assessed u images classified as "bad" were excluded ("After QC"). "After DM filter" and "After QC" are reported as the number of images (number of patients). Patient-lev available for IDRiD, as patient identifiers were not provided. Age (mean ± SD) and sex (number of male images, %) were summarised at the image level after IDRiD images were from patients with diabetes mellitus, so no DM filter was applied, and age and sex were not available for this dataset.

Abbreviations: BES, Beijing Eye Study; CIEMS, Central India Eye and Medical Study; DM, diabetes mellitus; ICDR, International Clinical Diabetic Retinopathy; Q standard deviation; UEMS, Ural Eye and Medical Study.

Evaluation of Predictive Performance for DR Screening

All three model architectures (ConvNeXt-Small,[23] ViT-Small,[24] and RETFound DINOv2[25]) demonstrated robust performance in detecting referable DR, with AUROCs ranging 0.960–0.974 across all three groups and three models. Relative to the internally tuned threshold, group-specific thresholds improved specificity, positive predictive value (PPV), and F1 score across all three groups and three models. Specificity ranged 0.699–0.869 with the internally tuned threshold and 0.900–0.935 with group-specific thresholds; PPV ranged 0.233–0.737 and 0.418–0.914, respectively; and F1 score 0.375–0.843 and 0.580–0.907. Notably, the population-based group showed substantially lower PPV and F1 score than the clinic-based group: across the three models and both threshold strategies, PPV ranged 0.233–0.459 versus 0.715–0.914, and F1 score ranged 0.375–0.608 versus 0.829–0.907.

Figure 2 illustrates model-level performance across the three external validation groups (all-datasets, clinic-based, and population-based). Detailed metrics, including AUROC, sensitivity, specificity, PPV, negative predictive value (NPV), and F1 score with 95% confidence intervals (CIs), are provided in Supplemental Table 1. Results for each internal and external validation dataset and model architecture with 95% CIs are shown in Supplemental Figs. 1 and 2.

Assessment of Probability Calibration and Uncertainty Alignment

Probability calibration showed substantial variation across the three external validation groups (all-datasets, clinic-based, and population-based) and three models, with the expected calibration error (ECE) ranging 0.031–0.133, the Brier score 0.056–0.081, and the negative log-likelihood (NLL) 0.219–0.278. Prevalence-based odds adjustment improved calibration across all groups and models, narrowing the ranges to 0.025–0.059 (ECE), 0.042–0.075 (Brier score), and 0.147–0.248 (NLL). ECE, Brier score, and NLL values with 95% CIs for the three external validation groups are shown in Fig. 3. Calibration plots and the corresponding metrics for the individual external datasets are provided in Supplemental Fig. 3.

Pearson, Spearman, and Kendall tau correlation analyses showed a modest association between model-derived (Monte Carlo [MC] dropout-based) and clinician-derived uncertainty. Across the two external validation datasets (HMADR_Clinic and HMADR_Screening) and three models, correlation coefficients

ranging 0.156–0.548 (Pearson), 0.151–0.532 (Spearman), and 0.122–0.424 (Kendall tau); all p-values were < 0.001. Violin plots in Fig. 4, stratified by clinician uncertainty quartiles, showed that model uncertainty remained consistently high even in the lowest quartile (Q1), suggesting limited alignment between model confidence and clinician-perceived uncertainty. Furthermore, analysis of model-derived uncertainty based on deep ensembles yielded results consistent with those obtained using MC dropout. Detailed results are provided in Supplemental Fig. 4.

Model Robustness to Noise and Cross-Device Variation

For intra-device pairs, all models demonstrated robust agreement overall, with intraclass correlation coefficient (ICC) values ranging 0.940–0.962 and Cohen's kappa values ranging 0.776–0.852. In contrast, agreement deteriorated markedly in the borderline subset, defined as image pairs in which at least one acquisition yielded a predicted probability close to the decision threshold, reflecting greater clinical and algorithmic uncertainty. Within this subset, ICC values ranged 0.051–0.220, and Cohen's kappa values ranged 0.141–0.331 across the three models.

Agreement between Triton and Maestro2 (both manufactured by Topcon) was consistently higher than comparisons involving Eidon. Across the three models, ICC values ranged from 0.920 to 0.951 for Triton–Maestro2, 0.678 to 0.772 for Triton–Eidon, and 0.686 to 0.760 for Maestro2–Eidon, with corresponding Cohen's kappa values ranging 0.686–0.795, 0.443–0.608, and 0.431–0.515, respectively. In borderline cases, agreement was uniformly poor across all device pairs and models, with ICC values ranging 0.024–0.136 and Cohen's kappa values ranging − 0.027–0.270. These results, together with their 95% CIs, are shown in Fig. 5, and a sensitivity analysis of borderline cases is provided in Supplemental Fig. 5.

Impact of Model Updates on Prediction Stability and Performance

All three model series showed consistently high AUROC and inter-update agreement in the three external validation groups (all-datasets, clinic-based, and population-based). Across the three models, three groups, and six inter-update comparisons, ICC values ranged 0.725–0.998 and Cohen's kappa values ranged 0.455–0.989 with the internally tuned threshold, and 0.770–0.997 (ICC) and 0.645–0.986 (kappa) with the group-specific optimal thresholds. In contrast, agreement declined in borderline cases. With the internally tuned threshold, ICC values ranged 0.063–0.958 (mean, 0.664) and Cohen's kappa values ranged 0.060–0.910 (mean, 0.463); with the group-specific optimal thresholds, ICC values ranged 0.101–0.943 (mean, 0.614) and Cohen's kappa values ranged 0.190–0.878 (mean, 0.582). The Youden-optimal thresholds (the value maximising sensitivity + specificity − 1) also exhibited notable variability across model versions, ranging 0.253–0.699 for all-datasets, 0.394–0.799 for clinic-based, and 0.107–0.669 for population-based.

Average inter-update agreement and trends in threshold instability for each group and model are illustrated in Fig. 6, with detailed results provided in Supplemental Fig. 6. A sensitivity analysis of borderline cases across model update scenarios is provided in Supplemental Fig. 7. To further assess the robustness of these findings, we retrained the models using five different random orderings of dataset updates. The overall patterns were consistent with the main results, as illustrated in Supplemental Fig. 8.

Discussion

In this study, we developed DR screening models and systematically evaluated multiple factors that are critical for real-world deployment, including probability calibration, uncertainty estimation, robustness to acquisition noise and cross-device variation, and consistency across model updates. Our findings suggest that, despite achieving high predictive performance, the models exhibited notable limitations in probability calibration, uncertainty estimation, robustness, and update consistency.

A meta-analysis of IDx-DR, the first FDA-approved product for DR screening, reported a pooled sensitivity of 0.95 and specificity of 0.91 across 13 studies.[26] In addition, a large-scale multi-centre DR screening study evaluating seven algorithms, including one FDA-approved product, from five companies across 311,604 images reported sensitivities ranging 0.510–0.860 and specificities 0.604–0.837.[27] Our models achieved strong discriminative performance, with a median AUROC of 0.979 (0.948–0.997) across three model architectures and four internal validation datasets, and 0.974 (0.904–0.993) across six external validation datasets, broadly consistent with performance levels previously reported for FDA-cleared and commercially deployed DR screening systems.[26,27] Detailed sensitivity and specificity values are summarised in Supplemental Table 1. However, PPV varied substantially across datasets, reflecting differences in disease prevalence. Across individual datasets and threshold strategies (Supplemental Fig. 2), PPV ranged 0.608–0.986 in eye clinic-based datasets (HMADR_Clinic and IDRiD), whereas in population-based datasets (UEMS, CIEMS, BES, and HMADR_Screening), it ranged 0.116–0.661.

These findings confirm that high sensitivity, specificity, and AUROC do not necessarily translate into high PPV, which is heavily dependent on disease prevalence in the target population. This observation aligns with concerns raised in prior multi-centre evaluations.[27] In population-based DR screening contexts where DR prevalence is relatively low, even algorithms with acceptable sensitivity and specificity may yield low PPV, resulting in a high proportion of false positives.[28,29] Such false positives can lead to unnecessary referrals, increased workload for human graders, and inefficient allocation of limited clinical resources.[30] This issue is closely tied to the threshold that converts model outputs into binary decisions (e.g., referable versus non-referable), and adjusting it might appear to offer a temporary solution. However, because these thresholds are locked at the time of regulatory approval, they cannot be readily adjusted in response to local prevalence without regulatory re-evaluation. Accordingly, our results therefore suggest that expected disease prevalence in the intended screening population should be explicitly incorporated into both development and validation frameworks.[31]

Experiments on probability calibration showed that distributional shifts between development and validation datasets can substantially degrade calibration, even when predictive performance remains high.[16,17] By applying a simple prevalence-based adjustment, calibration metrics, including ECE, Brier score, and NLL, improved in all groups, particularly population-based datasets, suggesting that prevalence shift is a meaningful contributor to miscalibration.[31] However,

the gains were modest in the clinic-based group, indicating that prevalence alone does not fully account for the observed miscalibration.[17] This is unsurprising given the substantial heterogeneity across our external datasets in terms of geography, ethnicity, age distribution, comorbidity profile, and imaging acquisition – factors that may alter the underlying relationship between retinal morphology and disease probability in ways that simple prevalence-based recalibration cannot correct. More broadly, when the deployment population differs from the training population, predicted probabilities may become systematically over- or under-estimated relative to the true event rate.[16,17,31] For example, a predicted probability of 0.8 may carry a very different clinical meaning when applied to an elderly rural Indian population than to a younger urban Korean screening cohort, even if the nominal DR prevalence is similar. Thus, even when sensitivity and specificity remain unchanged, mis-calibrated probabilities may lead to inappropriate reliance on, or dismissal of, model outputs, thereby limiting clinical utility.[32]

Our uncertainty analyses further showed that model-derived uncertainty was only weakly aligned with clinician-derived uncertainty. Although a modest correlation was observed, violin plots revealed that the model uncertainty and clinician disagreement may represent fundamentally different sources of ambiguity. This discrepancy is not unexpected, as deep learning models generate probabilistic outputs that do not necessarily correspond to true outcome likelihoods.[33] While Ayhan et al.[14] reported that higher model uncertainty in DR detection was associated with greater clinician disagreement under a test-time augmentation framework, our findings suggest that, although a statistical association exists, model uncertainty cannot be interpreted as a direct proxy for clinician-perceived diagnostic uncertainty.[15]

However, uncertainty estimation may still hold considerable promise for future clinical AI workflows. As AI becomes increasingly integrated into clinical workflows, uncertainty estimation may play a key role in screening settings such as primary care and optometry.[34,35] By identifying ambiguous cases for review by a trained physician or optometrist prior to specialist referral, the system could enable more efficient allocation of ophthalmic resources.[20,36]

To evaluate robustness to within-session acquisition noise and inter-device variability, we analysed prediction agreement across paired fundus photographs acquired from the same eye at the same visit. For intra-device pairs, overall agreement was high. Agreement was likewise high for Triton–Maestro2 pairs, but was lower for pairs involving the Eidon, likely reflecting its wider field of view relative to the Topcon devices. However, agreement declined markedly in borderline cases across all conditions. Prior multi-camera DR studies reported no significant population-level impact of fundus camera type on aggregate AUROC, a finding consistent with our overall results.[7–9,37] However, these aggregate comparisons obscure substantial instability at the individual image level, particularly for cases near the referral threshold –precisely those for whom the referral decision is most uncertain and most consequential for both the patient and the clinical pathway.

Similarly, prediction consistency across sequential model updates remained high at the aggregate level yet declined in borderline cases, even when the analysis was restricted to images correctly classified by the preceding model version. This pattern mirrors the robustness findings above: although overall performance on population-level metrics such as AUROC, remain stable individual-level referral decisions fluctuate following model updates.[38,39] The Youden-optimal threshold also exhibited notable variability across model versions. Few ophthalmic screening studies have explicitly examined decision retention across successive model versions, and current evaluation practices place greater emphasis on discrimination and cross-cohort generalisability than on update-induced decision drift.[1,6] Our findings suggest that update consistency, assessed at the individual rather than population level, and with particular attention to borderline cases, is a critical yet underexplored factor in evaluating models undergoing iterative refinement in clinical deployment.

In practice, the instability we observed across robustness and update consistency analyses has direct clinical consequences. A patient may be reclassified from "non-referable" to "referable", or vice versa, simply due to a model update, imaging device change, or acquisition variability, with no change in their underlying clinical status. It means that even small shifts in borderline decisions can substantially alter referral volumes, downstream workloads, and resource allocation across screening programmes. They can also undermine the longitudinal consistency that clinicians and patients reasonably expect from a deployed diagnostic tool. These observations carry concrete implications for real-world deployment in an era of growing emphasis on predetermined change-control plans and post-market surveillance. Developers should routinely report reliability dimensions beyond discrimination, including robustness, and pre/post-update agreement. Regulators, in turn, should treat model updates as events rather than routine refinements. While these updates may improve performance, they can also produce unintended consequences that are obscured by aggregate predictive metrics.

Several limitations warrant consideration. First, our models were developed using publicly available architectures and datasets; commercial systems may employ proprietary optimisation strategies that could affect the generalisability of our findings to deployed products. Second, despite integrating seven open-source datasets standardised using the International Clinical Diabetic Retinopathy (ICDR) scale, residual heterogeneity and label noise across sources may persist. Third, substantial variation in disease prevalence across external datasets limits the interpretability of direct cross-site metric comparisons. Fourth, update consistency was evaluated within a controlled, sequential framework that involved only incremental additions of geographically and institutionally distinct datasets. This design simulates a common real-world scenario in which models are progressively expanded to cover new sites, but does not capture temporal update scenarios in which models are retrained on longitudinally accumulated data from the same site. As such, the observed prediction stability may also underestimate the variability encountered in real-world deployment settings, where more substantial modifications to model architecture, optimization strategies, or loss functions may occur. Moreover, this evaluation framework is most relevant to continuously updated product-level systems under regulatory oversight, and may be less applicable to static cross-sectional research settings that are not iteratively revised.

In conclusion, our findings demonstrate that high predictive performance alone is insufficient for reliable AI-based screening. Each of the four dimensions in our experiments revealed systematic vulnerabilities invisible to conventional metrics, with direct consequences for clinical decision-making. These dimensions should be standard components of AI evaluation, aligning with emerging reporting frameworks such as TRIPOD+AI[40] and evolving post-market regulatory guidance. In parallel, our results emphasise that development and validation datasets should reflect the intended deployment context:[31] a model designed for population-level screening is not interchangeable with one designed for tertiary practice, and treating them as such obscures clinically

meaningful differences. To facilitate further research on uncertainty estimation in DR screening, we also publicly release the HMADR_Clinic dataset, independently annotated by five ophthalmologists and adjudicated by a senior retinal specialist.

Methods

Task definitions

We aim to evaluate a DR screening tool for binary classification of referable versus non-referable disease. In accordance with the ICDR severity scale,[41] referable DR is defined as moderate non-proliferative DR (NPDR), severe NPDR, or proliferative DR (PDR) (ICDR grades 2–4), consistent with established clinical referral thresholds and prior large-scale screening studies.[19,42] Non-referable DR comprises cases graded as no DR (normal) or mild NPDR (ICDR grades 0–1).

Development Datasets

The development dataset comprised seven publicly available fundus image datasets: EyePACS (United States),[43] BRSET (Brazil),[44] mBRSET (Brazil),[45] APTOS 2019 (India),[46] DeepDRiD (China),[47] DDR(China),[48] and MESSIDOR2 (France).[49] These datasets were selected to maximise training volume, standardise image acquisition, ensure consistent annotation using the five-level ICDR grading scale, and leverage publicly available resources that have been validated in prior studies and challenges. In particular, APTOS 2019 and MESSIDOR2 have been widely adopted as benchmark datasets in the RETFound model series. Although mBRSET was acquired with a portable smartphone-based handheld fundus camera rather than a conventional tabletop device, it was included because its 45-degree field of view is comparable to that of standard fundus cameras. DeepDRiD includes two types of fundus photographs: conventional (e.g., 45°) and ultra-wide field images. We excluded the ultra-wide images, and DeepDRiD in this paper refers exclusively to the conventional fundus photographs. Detailed dataset characteristics, including country of origin, acquisition device, fixation protocol, number of patients and images, and the resulting split sizes after quality control, are summarised in Table 1.

External Validation Datasets

For external validation, six independent datasets representing two clinically distinct settings were used. The first category comprised datasets from eye clinics: IDRiD (India), a publicly available challenge dataset for DR grading; and HMADR_Clinic (South Korea), a newly collected dataset comprising fundus images acquired at the ophthalmology outpatient clinic of Hanyang University Guri Hospital, a tertiary medical centre, between March 1, 2020 and March 1, 2024, restricted to patients with available HbA1c measurements.

The second category represented population-level screening and epidemiological settings. This category included three established population-based epidemiological studies —UEMS (Russia),[50] BES (China),[51] and CIEMS (India)[52] — each providing macula- and disc-centred fundus images. In addition, HMADR_Screening (South Korea) is a newly collected dataset of fundus images obtained at the affiliated health screening centre of Hanyang University Guri Hospital over the same period as HMADR_Clinic (March 1, 2020 – March 1, 2024), restricted to participants with available HbA1c measurements, and represents an asymptomatic general screening population. Consistent with the intended use of DR screening tools, analyses were restricted to participants with diabetes mellitus (HbA1c ≥ 6.5%); all IDRiD images were already from patients with diabetes mellitus and required no further filtering. Detailed dataset characteristics, including country, acquisition device, fixation protocol, number of patients and images, demographic summaries, and image counts after DM filtering and quality control, are summarised in Table 2.

Datasets for Robustness to Noise and Cross-Device Variation

The AI-READI dataset[53] comprises multimodal retinal imaging data from 2,280 participants, including patients with type 2 diabetes mellitus and healthy controls, collected across multiple sites in the United States using several imaging systems, including the Topcon Triton, Topcon Maestro2 (Topcon Corporation, Tokyo, Japan), iCARE Eidon (iCare Finland Oy, Vantaa, Finland), and Optomed Aurora IQ (Optomed Oyj, Oulu, Finland). Notably, the Topcon Triton optical coherence tomography angiography (OCTA) system simultaneously generates a fundus image during both 6 × 6 mm and 12 × 12 mm scan acquisitions.

Institutional Data Collection and Ethics

Collection and analysis of the HMADR_Clinic and HMADR_Screening datasets were approved by the Institutional Review Board of Hanyang University Guri Hospital (IRB No. 2024-12-007-001). The study was conducted in accordance with the tenets of the Declaration of Helsinki, and the requirement for informed consent was waived owing to the retrospective nature of the study.

Image Grading and Annotation Protocol

For UEMS, BES, and CIEMS—population-based epidemiological studies—fundus images were systematically annotated by ophthalmologists with over 20 years of clinical experience using the five-level ICDR grading scale. Annotation proceeded in two stages. First, individual DR features were labelled, including haemorrhages, microaneurysms, hard and soft exudates, intraretinal microvascular abnormalities, venous beading, neovascularisation at the optic disc and elsewhere, proliferative fibrous tissue, macular oedema, and photocoagulation scars. Subsequently, the ICDR grade was derived from the combination of these lesion-level annotations.

For the HMADR_Clinic and HMADR_Screening datasets, five board-certified ophthalmologists, each with more than 10 years of experience in retinal diseases, independently graded all images using a five-level ICDR severity scale. Each grader independently assigned a single grade per image via a standardised interface presenting one image at a time, without access to other graders' assessments. For each image, the final label was assigned by majority vote among the five annotators, with the senior retinal specialist's label used as the tie-breaker when two or more classes received an equal number of votes. The

resulting five-level ICDR grade was then binarised into referable (grades 2–4) and non-referable (grades 0–1) categories. Detailed distributions of referable and non-referable cases are presented in Table 2, and detailed annotation results with agreement analysis, including Cohen's kappa and Gwet's AC1,[54] for HMADR_Clinic and HMADR_Screening are provided in Supplemental Table 2.

Dataset Preprocessing

Image quality control for BRSET, mBRSET, and DeepDRiD was performed using dataset-provided quality labels, which are released as binarised labels: the final_quality label for BRSET, the quality label for mBRSET, and the Overall quality label for DeepDRiD. For all other datasets, including external validation sets and AI-READI, automated quality assessment was conducted using AutoMorph, which classifies fundus images as good, usable, or bad; only images classified as good or usable were retained. All included images were resized to 512× 512 pixels. Across all datasets, five-level ICDR labels were binarised into non-referable (grades 0–1) and referable (grades 2–4) categories.

Among the internal datasets, APTOS 2019, MESSIDOR2, and DDR were used exclusively for training, since these datasets lack the patient-level identifiers required for patient-wise splitting. The remaining datasets—EyePACS, BRSET, mBRSET, and DeepDRiD —were split into training, tuning, and validation sets at a 6:2:2 ratio, stratified by the presence of referable DR. To prevent data leakage, all splits were performed at the patient level.

For all external validation datasets, only fundus images from patients with diabetes mellitus were included, consistent with the intended clinical application of DR screening. In all datasets except IDRiD, diabetes mellitus was defined as an HbA1c level of 6.5% or higher, in accordance with the American Diabetes Association criteria. The IDRiD dataset was used in its entirety, as it was originally curated to include only patients with diabetes mellitus. In the BES—a longitudinal epidemiological study with fundus photography conducted in 2001, 2006, and 2011—only images acquired at the 2011 examination were included. Detailed case counts following quality and diabetes mellitus filtering are provided in Tables 1 and 2.

Model development and incremental training design

Models were implemented in PyTorch, with data augmentation performed using Albumentations and pretrained backbones obtained from timm. Three architectures were evaluated: a DINOv3[55]–pretrained ConvNeXt-small, a DINOv3–pretrained Vision Transformer (ViT)-small, and RETFound DINOv2, a foundation model for fundus photography pretrained on large-scale retinal image data using DINOv2.

For each architecture, seven model versions were trained sequentially by incrementally adding one internal dataset at a time in the following order: MESSIDOR2 (Model 1), EyePACS (Model 2), APTOS 2019 (Model 3), DDR (Model 4), DeepDRiD (Model 5), BRSET (Model 6), and mBRSET (Model 7). The order was determined by the chronological release date of each dataset. This incremental design was intended to simulate the progressive accumulation of training data during model development and maintenance, and to enable systematic evaluation of how successive updates influence decision stability.

All models were trained for 50 epochs using the AdamW[56] optimiser with a learning rate of 2×10^{-4}. Mixed-precision training was employed to reduce memory usage and improve computational efficiency. Cross-entropy loss was used as the objective function. Data augmentation strategies included RandomResizedCrop (scale 0.8–1.0), HorizontalFlip, and RandomRotate90, as well as affine transformations (shear and rotation), and ColorJitter to enhance robustness and generalisation. A fixed random seed (42) was used across all experiments, and PyTorch was configured in deterministic mode to ensure reproducibility. This ensured identical initialisation of the classification head weights across architectures, thereby isolating the effect of incremental data addition on model behaviour.

To assess the robustness of our findings, we additionally trained models under two modified configurations: (1) using five random seeds (42, 43, 44, 45, 46) to enable deep ensemble–based uncertainty estimation; and (2) using 5 different random dataset orderings to assess the stability of model update patterns. The five orderings were as follows: (1) EyePACS, DDR, DeepDRiD, APTOS 2019, mBRSET, Messidor2, BRSET; (2) DDR, mBRSET, APTOS 2019, DeepDRiD, Messidor2, BRSET, EyePACS; (3) DDR, APTOS 2019, DeepDRiD, EyePACS, BRSET, mBRSET, Messidor2; (4) APTOS 2019, BRSET, Messidor2, DDR, mBRSET, EyePACS, DeepDRiD; and (5) mBRSET, DDR, BRSET, EyePACS, Messidor2, APTOS 2019, DeepDRiD.

Predictive performance metrics

The models' predictive performance for referable DR was evaluated using AUROC, sensitivity, specificity, PPV, NPV, and F1 score. Two decision thresholds, determined by maximising the Youden Index, were applied: (i) an internally tuned threshold derived from the internal tuning dataset, and (ii) group-specific thresholds independently optimised for each external cohort. Performance was assessed across three aggregated external validation settings: (1) all six external datasets combined (all-datasets group); (2) two clinic-based datasets (IDRiD and HMADR_Clinic); and (3) four population-based datasets (UEMS, BES, CIEMS, and HMADR_Screening).

Probability calibration

Model calibration was assessed using decile-based ECE, Brier score, NLL, and calibration plots. Predicted probabilities were partitioned into 10 equal-frequency bins (deciles), within which the mean predicted probability and observed event frequency were computed. ECE was defined as the average absolute difference between these quantities across all bins. The Brier score measures the mean squared error between predicted probabilities and observed binary outcomes, capturing both calibration and sharpness. NLL quantifies the log-loss associated with the predicted probability assigned to the true class, penalising overconfident incorrect predictions more heavily. Calibration plots were constructed by plotting the mean predicted probability against the observed event frequency for each bin.

To examine whether the observed miscalibration reflected prevalence shift rather than intrinsic model limitations, we applied a simple post-hoc rescaling of predicted probabilities, adapting from the prevalence-correction framework of Saerens et al.[57] Specifically, predicted probabilities were rescaled in odds space so that their mean on each target dataset matched the empirical prevalence of that dataset. This adjustment requires no model retraining or fitting of

additional parameters. Models were subsequently re-evaluated using ECE, Brier score, NLL, and calibration plots to quantify the impact of this adjustment. Because this adjustment depends only on the target prevalence and leaves the model itself unchanged, improvements in calibration metrics can be interpreted as evidence that a prevalence shift between the development and deployment populations is a contributing factor.

Uncertainty estimation

Model uncertainty was estimated using MC dropout, whereby dropout layers are retained at test time to enable stochastic inference. This approach assumes that confident models yield low variance across repeated predictions, whereas uncertain models exhibit greater variability. For each image, uncertainty was quantified as the standard deviation of predicted probabilities obtained from 50 stochastic forward passes. For the two HMADR datasets, clinician-derived uncertainty was estimated as the inter-grader variability among five independent board-certified ophthalmologists. The sixth senior ophthalmologist was excluded from this analysis, as this grader's role was limited to adjudicating cases of disagreement and did not include providing independent grades for all images. Analogous to the model-based approach, this was quantified as the standard deviation of the binary labels assigned by the five ophthalmologists for each image. Agreement between model and clinician uncertainty was assessed by using Pearson's correlation, Spearman's rank correlation, and Kendall tau correlation, along with their corresponding p-values, between per-image model standard deviation and inter-annotator label standard deviation. This relationship was further examined using violin plots across all datasets and model architectures. To corroborate these findings, model uncertainty was additionally estimated using deep ensembles, with detailed results provided in Supplemental Fig. 4.

Robustness evaluation

Robustness to acquisition noise and inter-device variability was evaluated by quantifying prediction agreement across paired fundus image acquisitions. Noise robustness was assessed via intra-device repeatability using paired images from the AI-READI dataset, in which each eye was imaged twice with the Triton device (6 × 6 and 12 × 12 OCTA scan protocols), yielding two fundus images per eye. Inter-device agreement was assessed across three cross-device pairs: Triton versus Maestro2, Triton versus Eidon, and Maestro2 versus Eidon. The Aurora IQ was excluded due to inconsistent fixation across acquisitions. Agreement was quantified using Cohen's kappa for categorical label and the ICC for continuous probabilities. Analyses were conducted on both the full set of paired images and on a borderline subset, defined as pairs in which at least one image had a predicted probability near the decision threshold. The borderline subset was defined as the 10% of samples closest to the decision threshold on each side (i.e., the closest 10% above and the closest 10% below). To evaluate the robustness of this definition, a sensitivity analysis varying the borderline fraction from 10% to 90% was performed and is provided in Supplemental Fig. 5.

Update consistency analysis

To assess whether successive model updates preserved prior decisions, prediction consistency was evaluated across sequential model versions. For each update pair (model *i* -> model *i* + 1), agreement was computed exclusively on images correctly classified by model *i*, thereby quantifying the extent to which model *i* + 1 retained previously correct decisions. This design isolates update-induced decision drift from pre-existing misclassifications. Agreement was quantified using Cohen's kappa and ICC, computed separately for the full datasets and the borderline subset defined above. AUROC was additionally calculated for each model version to track overall performance across updates. The operational stability of Youden-optimal thresholds was also evaluated across datasets and model updates. In the main analysis, model updates followed the chronological release order of the datasets. To confirm the robustness of these findings, we additionally performed two analyses. First, a sensitivity analysis varying the borderline fraction from 10% to 90% in 10% increments was conducted, and the results are presented in Supplemental Fig. 7. Second, models were retrained under five random update orderings, with the corresponding results provided in Supplemental Fig. 8.

Statistical analysis

For all metrics, 95% CIs were estimated via bootstrap resampling (1,000 iterations, with replacement from the original sample) and reported as the 2.5th and 97.5th percentiles of the bootstrap distribution. Unless otherwise specified, values in parentheses represent the range (minimum–maximum) across models, datasets, or threshold strategies. ICCs were calculated using a two-way random-effects model for absolute agreement of single measurements (ICC2). All analyses were conducted in Python using *scikit-learn*, *SciPy*, and *pingouin*.

Data and code availability

The publicly available datasets used in this study can be accessed as follows: EyePACS (https://www.kaggle.com/c/diabetic-retinopathy-detection), BRSET (https://physionet.org/content/brazilian-ophthalmological/), mBRSET (https://physionet.org/content/mbrset/), APTOS 2019 (https://www.kaggle.com/c/aptos2019-blindness-detection), MESSIDOR2 (https://www.adcis.net/en/third-party/messidor2/), DeepDRiD (https://github.com/deepdrdoc/DeepDRiD), DDR (https://github.com/nkicsl/DDR-dataset), and IDRiD (https://idrid.grand-challenge.org/). The UEMS, CIEMS, and BES datasets are not publicly available due to participant privacy and institutional data-sharing agreements, but de-identified data may be made available from the corresponding investigators (M.M.B./G.M.K., V.N., and J.B.J./Y.W., respectively) upon reasonable request and approval by the relevant institutional review boards.

The HMADR_Clinic dataset, including multi-annotator labels, is publicly available at https://github.com/hyunminkim0625/HMADR_Clinic_dataset under a CC BY-NC 4.0 license. The HMADR_Screening dataset is available from the corresponding author (S.H.) upon reasonable request, subject to Hanyang University IRB approval and a data-use agreement.

The code used for model training and evaluation was adapted from the RETFound repository (https://github.com/rmaphoh/RETFound) and is implemented in PyTorch. Pretrained weights for ConvNeXt-Small (DINOv3) and ViT-Small (DINOv3) are available at https://github.com/facebookresearch/dinov3, and RETFound DINOv2 weights are available at https://huggingface.co/YukunZhou. Image quality control was performed using the automated retinal image analysis tool AutoMorph (https://github.com/rmaphoh/AutoMorph).

Declarations

Competing Interests

P.A.K. is a co-founder of Cascader Ltd. and has acted as a consultant for Skleo Health, insitro, Retina Consultants of America, Roche, Boehringer-Ingelheim, and Bitfount, and is an equity owner in Big Picture Medical. He has received speaker fees from Zeiss, Théa, Apellis, and Roche, travel support from Bayer and Roche, and grant funding from Roche. He has attended advisory boards for Topcon, Bayer, Boehringer-Ingelheim, and Roche.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contribution

H.K. and R.A.J. designed the study and implemented the training and evaluation pipeline and performed statistical analyses. T.N., R.H., V.S., A.H., R.A.J., and S.H. performed annotation. A.Y.O., M.Gr., B.P., and Y.Z. contributed to study conceptualization. L.J., Y.L., E.R., D.W., J.Hu., K.I., P.N., M.G.M., M.E., T.E., and E.F.T. contributed to data preprocessing, model development, validation, and manuscript writing. J.C. and D.L. collected and managed datasets. Y.W., S.P.J., J.B.J., M.M.B., G.M.K., and V.N. managed cohort data and provided clinical expertise. S.K.W., S.H., and P.A.K. supervised the project. All authors contributed to manuscript writing and approved the final version.

Acknowledgement

The research was supported by the National Institute for Health and Care Research (NIHR) Biomedical Research Centre based at Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Data Availability

The publicly available datasets used in this study can be accessed as follows: EyePACS (https://www.kaggle.com/c/diabetic-retinopathy-detection), BRSET (https://physionet.org/content/brazilian-ophthalmological/), mBRSET (https://physionet.org/content/mbrset/), APTOS-2019 (https://www.kaggle.com/c/aptos2019-blindness-detection), MESSIDOR2 (https://www.adcis.net/en/third-party/messidor2/), DeepDRiD (https://github.com/deepdrdoc/DeepDRiD), DDR (https://github.com/nkicsl/DDR-dataset), and IDRiD (https://idrid.grand-challenge.org/). The UEMS, CIEMS, and BES datasets are not publicly available due to participant privacy and institutional data-sharing agreements, but de-identified data may be made available from the corresponding investigators (M.M.B./G.M.K., V.N., and J.B.J./Y.W., respectively) upon reasonable request and approval by the relevant institutional review boards.The HMADR_Clinic dataset, including multi-annotator labels, is publicly available at https://github.com/hyunminkim0625/HMADR_Clinic_dataset under a CC BY-NC 4.0 license. The HMADR_Screening dataset is available from the corresponding author (S.H.) upon reasonable request, subject to Hanyang University IRB approval and a data-use agreement.The code used for model training and evaluation was adapted from the RETFound repository (https://github.com/rmaphoh/RETFound) and is implemented in PyTorch. Pretrained weights for ConvNeXt-Small (DINOv3) and ViT-Small (DINOv3) are available at https://github.com/facebookresearch/dinov3, and RETFound DINOv2 weights are available at https://huggingface.co/YukunZhou. Image quality control was performed using the automated retinal image analysis tool AutoMorph v.1.0 (https://github.com/rmaphoh/AutoMorph).

References

1. Muralidharan, V. *et al.* A scoping review of reporting gaps in FDA-approved AI medical devices. *NPJ Digit Med* 7, 273 (2024). https://doi.org/10.1038/s41746-024-01270-x
2. Wu, K. *et al.* Characterizing the clinical adoption of medical AI devices through US insurance claims. *Nejm Ai* 1, AIoa2300030 (2024).
3. Chouffani El Fassi, S. *et al.* Not all AI health tools with regulatory authorization are clinically validated. *Nat Med* 30, 2718–2720 (2024). https://doi.org/10.1038/s41591-024-03203-3
4. Wiens, J. *et al.* Do no harm: a roadmap for responsible machine learning for health care. *Nat Med* 25, 1337–1340 (2019). https://doi.org/10.1038/s41591-019-0548-6
5. Windecker, D. *et al.* Generalizability of FDA-Approved AI-Enabled Medical Devices for Clinical Use. *JAMA Netw Open* 8, e258052 (2025). https://doi.org/10.1001/jamanetworkopen.2025.8052
6. Andersen, E. S. *et al.* Monitoring performance of clinical artificial intelligence in health care: a scoping review. *JBI Evid Synth* 22, 2423–2446 (2024). https://doi.org/10.11124/JBIES-24-00042
7. Yip, M. Y. T. *et al.* Technical and imaging factors influencing performance of deep learning systems for diabetic retinopathy. *NPJ Digit Med* 3, 40 (2020). https://doi.org/10.1038/s41746-020-0247-1

8. Srinivasan, R., Surya, J., Ruamviboonsuk, P., Chotcomwongse, P. & Raman, R. Influence of Different Types of Retinal Cameras on the Performance of Deep Learning Algorithms in Diabetic Retinopathy Screening. *Life (Basel)* 12 (2022). https://doi.org/10.3390/life12101610
9. Tsai, M. J. *et al.* Cross-Camera External Validation for Artificial Intelligence Software in Diagnosis of Diabetic Retinopathy. *J Diabetes Res* 2022, 5779276 (2022). https://doi.org/10.1155/2022/5779276
10. Ong, A. Y. *et al.* Considering the missing science of retraining and maintenance in medical artificial intelligence, using ophthalmology as an exemplar. *NPJ Digit Med* 9 (2026). https://doi.org/10.1038/s41746-026-02463-2
11. Rosen, K. L. & Mandl, K. D. Regulating Flexibility for Artificial Intelligence FDA Experience with Predetermined Change Control Plans. *medRxiv*, 2025.2008.2026.25334477 (2025).
12. González, C. *et al.* Regulating radiology AI medical devices that evolve in their lifecycle. *arXiv preprint arXiv:2412.20498* (2024).
13. Lambert, B., Forbes, F., Doyle, S., Dehaene, H. & Dojat, M. Trustworthy clinical AI solutions: A unified review of uncertainty quantification in Deep Learning models for medical image analysis. *Artif Intell Med* 150, 102830 (2024). https://doi.org/10.1016/j.artmed.2024.102830
14. Ayhan, M. S. *et al.* Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection. *Med Image Anal* 64, 101724 (2020). https://doi.org/10.1016/j.media.2020.101724
15. Begoli, E., Bhattacharya, T. & Kusnezov, D. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence* 1, 20–23 (2019).
16. Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. in *International conference on machine learning.* 1321–1330 (PMLR).
17. Ovadia, Y. *et al.* Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems* 32 (2019).
18. Huang, Y., Li, W., Macheret, F., Gabriel, R. A. & Ohno-Machado, L. A tutorial on calibration measurements and calibration models for clinical prediction models. *J Am Med Inform Assoc* 27, 621–633 (2020). https://doi.org/10.1093/jamia/ocz228
19. Abramoff, M. D., Lavin, P. T., Birch, M., Shah, N. & Folk, J. C. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med* 1, 39 (2018). https://doi.org/10.1038/s41746-018-0040-6
20. Nderitu, P. & Keane, P. A. Real-World Adoption of Artificial Intelligence in Diabetic Retinopathy Screening: What Is the Current Status? *Clin Exp Ophthalmol* 53, 741–743 (2025). https://doi.org/10.1111/ceo.14605
21. Ruamviboonsuk, P. *et al.* Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. *NPJ digital medicine* 2, 25 (2019).
22. Macdonald, T. *et al.* Generating evidence to support the role of AI in diabetic eye screening: considerations from the UK National Screening Committee. *Lancet Digit Health* 7, 100840 (2025). https://doi.org/10.1016/j.landig.2024.12.004
23. Liu, Z. *et al.* in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 11976–11986.
24. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv*:2010.*11929* (2020).
25. Zhou, Y. *et al.* Generalist versus Specialist Vision Foundation Models for Ocular Disease and Oculomics. *arXiv preprint arXiv:2509.03421* (2025).
26. Khan, Z. *et al.* Diagnostic Accuracy of IDX-DR for Detecting Diabetic Retinopathy: A Systematic Review and Meta-Analysis. *Am J Ophthalmol* 273, 192–204 (2025). https://doi.org/10.1016/j.ajo.2025.02.022
27. Lee, A. Y. *et al.* Multicenter, Head-to-Head, Real-World Validation Study of Seven Automated Artificial Intelligence Diabetic Retinopathy Screening Systems. *Diabetes Care* 44, 1168–1175 (2021). https://doi.org/10.2337/dc20-1877
28. Wewetzer, L., Held, L. A. & Steinhauser, J. Diagnostic performance of deep-learning-based screening methods for diabetic retinopathy in primary care-A meta-analysis. *PLoS One* 16, e0255034 (2021). https://doi.org/10.1371/journal.pone.0255034
29. Kanagasingam, Y. *et al.* Evaluation of Artificial Intelligence-Based Grading of Diabetic Retinopathy in Primary Care. *JAMA Netw Open* 1, e182665 (2018). https://doi.org/10.1001/jamanetworkopen.2018.2665
30. Tufail, A. *et al.* Automated Diabetic Retinopathy Image Assessment Software: Diagnostic Accuracy and Cost-Effectiveness Compared with Human Graders. *Ophthalmology* 124, 343–351 (2017). https://doi.org/10.1016/j.ophtha.2016.11.014
31. Godau, P. *et al.* in *International Conference on Medical Image Computing and Computer-Assisted Intervention.* 389–399 (Springer).
32. Gaube, S. *et al.* Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ Digit Med* 4, 31 (2021). https://doi.org/10.1038/s41746-021-00385-9
33. Hüllermeier, E. & Waegeman, W. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning* 110, 457–506 (2021).
34. Wundram, A. M. & Baumgartner, C. F. in *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging.* 34–44 (Springer).
35. Leibig, C., Allken, V., Ayhan, M. S., Berens, P. & Wahl, S. Leveraging uncertainty information from deep neural networks for disease detection. *Sci Rep* 7, 17816 (2017). https://doi.org/10.1038/s41598-017-17876-z
36. Liu, J. *et al.* Diabetic Retinopathy Screening with Automated Retinal Image Analysis in a Primary Care Setting Improves Adherence to Ophthalmic Care. *Ophthalmol Retina* 5, 71–77 (2021). https://doi.org/10.1016/j.oret.2020.06.016
37. Ting, D. S. W. *et al.* Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. *JAMA* 318, 2211–2223 (2017). https://doi.org/10.1001/jama.2017.18152

38. Miller, E. W. & Blume, J. D. Diagnostics for Individual-Level Prediction Instability in Machine Learning for Healthcare. *arXiv preprint arXiv:2603.00192* (2026).
39. Cabanillas Silva, P. *et al.* Longitudinal Model Shifts of Machine Learning-Based Clinical Risk Prediction Models: Evaluation Study of Multiple Use Cases Across Different Hospitals. *J Med Internet Res* 26, e51409 (2024). https://doi.org/10.2196/51409
40. Collins, G. S. *et al.* TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 385, e078378 (2024). https://doi.org/10.1136/bmj-2023-078378
41. Wilkinson, C. P. *et al.* Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology* 110, 1677–1682 (2003). https://doi.org/10.1016/S0161-6420(03)00475-5
42. Gulshan, V. *et al.* Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* 316, 2402–2410 (2016). https://doi.org/10.1001/jama.2016.17216
43. Cuadros, J. & Bresnick, G. EyePACS: an adaptable telemedicine system for diabetic retinopathy screening. *J Diabetes Sci Technol* 3, 509–516 (2009). https://doi.org/10.1177/193229680900300315
44. Nakayama, L. F. *et al.* BRSET: A Brazilian Multilabel Ophthalmological Dataset of Retina Fundus Photos. *PLOS Digit Health* 3, e0000454 (2024). https://doi.org/10.1371/journal.pdig.0000454
45. Wu, C. *et al.* A portable retina fundus photos dataset for clinical, demographic, and diabetic retinopathy prediction. *Sci Data* 12, 323 (2025). https://doi.org/10.1038/s41597-025-04627-3
46. Karthik, M. & Dane, S. Aptos 2019 blindness detection. *Kaggle* https://kaggle.com/competitions/aptos2019-blindness-detection *Go to reference in* 5, 8 (2019).
47. Porwal, P. *et al.* IDRiD: Diabetic Retinopathy - Segmentation and Grading Challenge. *Med Image Anal* 59, 101561 (2020). https://doi.org/10.1016/j.media.2019.101561
48. Li, T. *et al.* Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Information Sciences* 501, 511–522 (2019).
49. Decencière, E. *et al.* Feedback on a publicly distributed image database: the Messidor database. *Image Analysis & Stereology*, 231–234 (2014).
50. Bikbov, M. M. *et al.* Prevalence and causes of vision impairment and blindness in the Russian ural eye and medical study. *Sci Rep* 10, 12397 (2020). https://doi.org/10.1038/s41598-020-69439-4
51. Jonas, J. B., Xu, L. & Wang, Y. X. The Beijing Eye Study. *Acta Ophthalmol* 87, 247–261 (2009). https://doi.org/10.1111/j.1755-3768.2008.01385.x
52. Nangia, V. *et al.* Ocular axial length and its associations in an adult population of central rural India: the Central India Eye and Medical Study. *Ophthalmology* 117, 1360–1366 (2010).
53. Owsley, C. *et al.* Cross-sectional design and protocol for artificial intelligence ready and equitable atlas for diabetes insights (AI-READI). *BMJ open* 15, e097449 (2025).
54. Gwet, K. Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Statistical methods for inter-rater reliability assessment* 1, 1–6 (2002).
55. Siméoni, O. *et al.* Dinov3. *arXiv preprint arXiv:2508.10104* (2025).
56. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).
57. Saerens, M., Latinne, P. & Decaestecker, C. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural Comput* 14, 21–41 (2002). https://doi.org/10.1162/089976602753284446

Figures

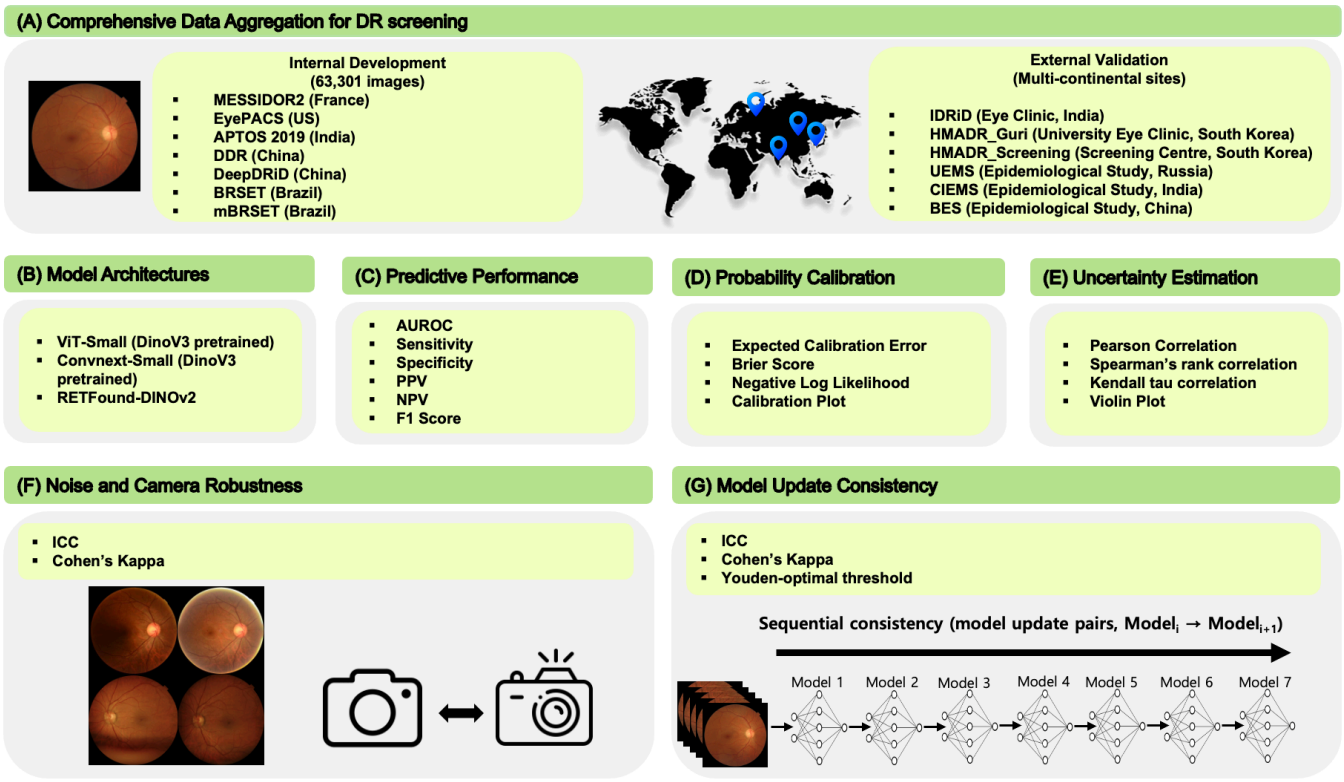


Figure 1

Study Overview

Overview of the study design, comprising seven analytical components. (a) Data aggregation: Seven internal development datasets (63,301 images) from the United States, Brazil, India, China, and France were used for model training and internal validation. In addition, six external validation datasets were obtained from multi-continental sites, including eye clinic, screening, and population settings. (b) Model architectures: Three architectures were evaluated: ViT-Small, ConvNeXt-Small, and RETFound DINOv2. (c) Predictive performance: Predictive performance was assessed using AUROC, sensitivity, specificity, PPV, NPV, and F1 score. (d) Probability calibration: Calibration was evaluated using expected calibration error, Brier score, negative log-likelihood, and calibration plots. (e) Uncertainty estimation: Model-derived and clinician-derived uncertainty were compared using Pearson's correlation, Spearman's rank correlation, Kendall tau correlation, and violin plots. (f) Noise and camera robustness: Prediction agreement across paired acquisitions and cross-device image pairs was quantified using ICC and Cohen's kappa. (g) Model update consistency: Prediction consistency across seven sequential model versions (v1–v7) was assessed for consecutive model pairs using ICC, Cohen's kappa, and group-specific Youden-optimal thresholds, and fluctuation of group-specific Youden-optimal thresholds.

Abbreviations: AUROC, area under the receiver operating characteristic curve; BES, Beijing Eye Study; CIEMS, Central India Eye and Medical Study; ICC, intraclass correlation coefficient; NPV, negative predictive value; PPV, positive predictive value; UEMS, Ural Eye and Medical Study.

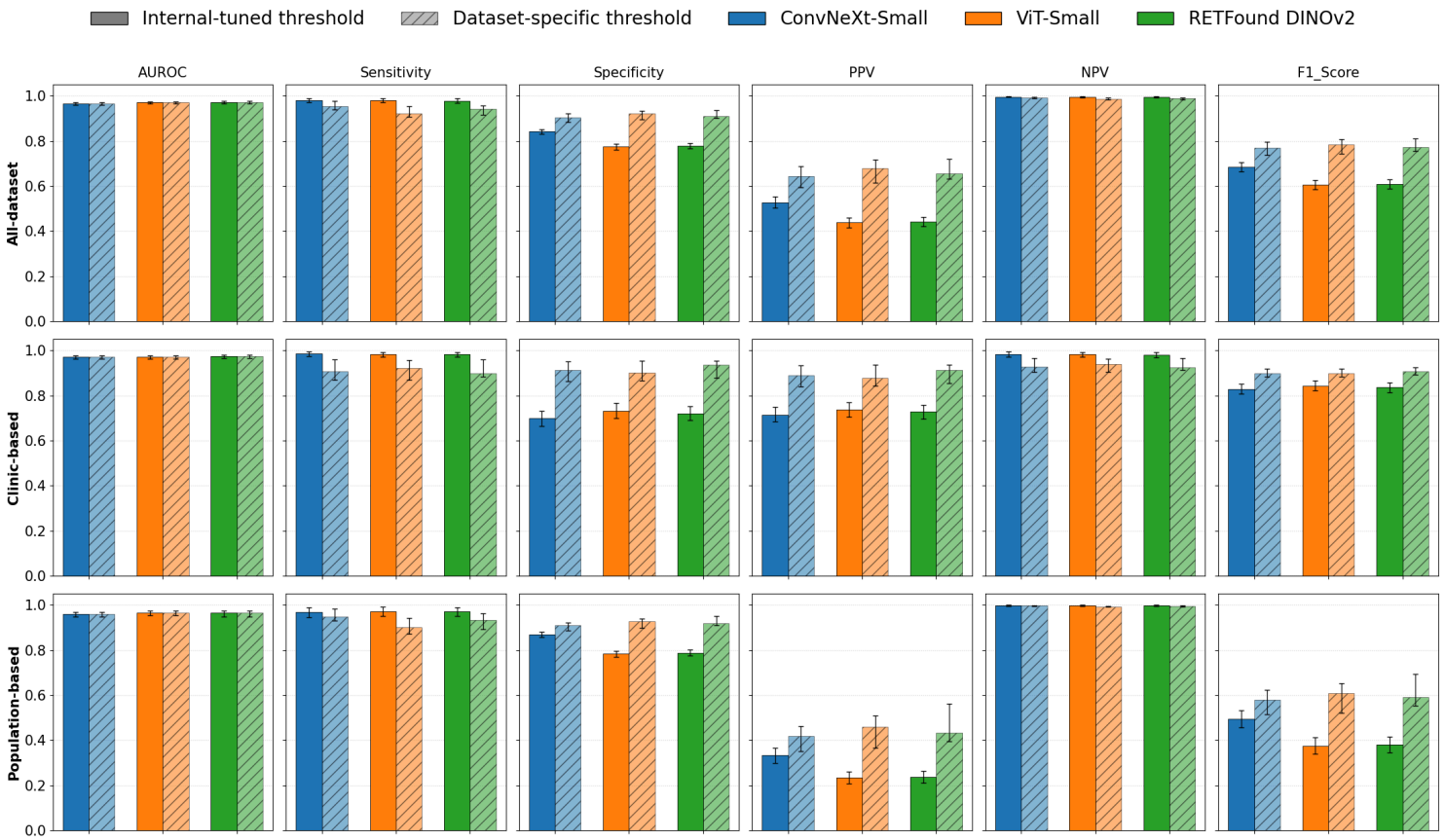


Figure 2

External validation discrimination performance across model architectures and datasets.

Bar charts summarise six performance metrics—AUROC, sensitivity, specificity, PPV, NPV, and F1 score—for three model architectures (ConvNeXt-Small, ViT-Small, and RETFound DINOv2) across three dataset groupings (all-datasets, clinic-based, population-based), evaluated under two threshold strategies (internally tuned and group-specific). Solid bars denote the internally tuned threshold, whereas hatched bars denote the group-specific threshold. AUROC, sensitivity, and NPV remained consistently high across models and settings. By contrast, PPV and F1 score exhibited marked heterogeneity, particularly in population-based datasets, due to the strong influence of disease prevalence.

Abbreviations: AUROC, area under the receiver operating characteristic curve; NPV, negative predictive value; PPV, positive predictive value.

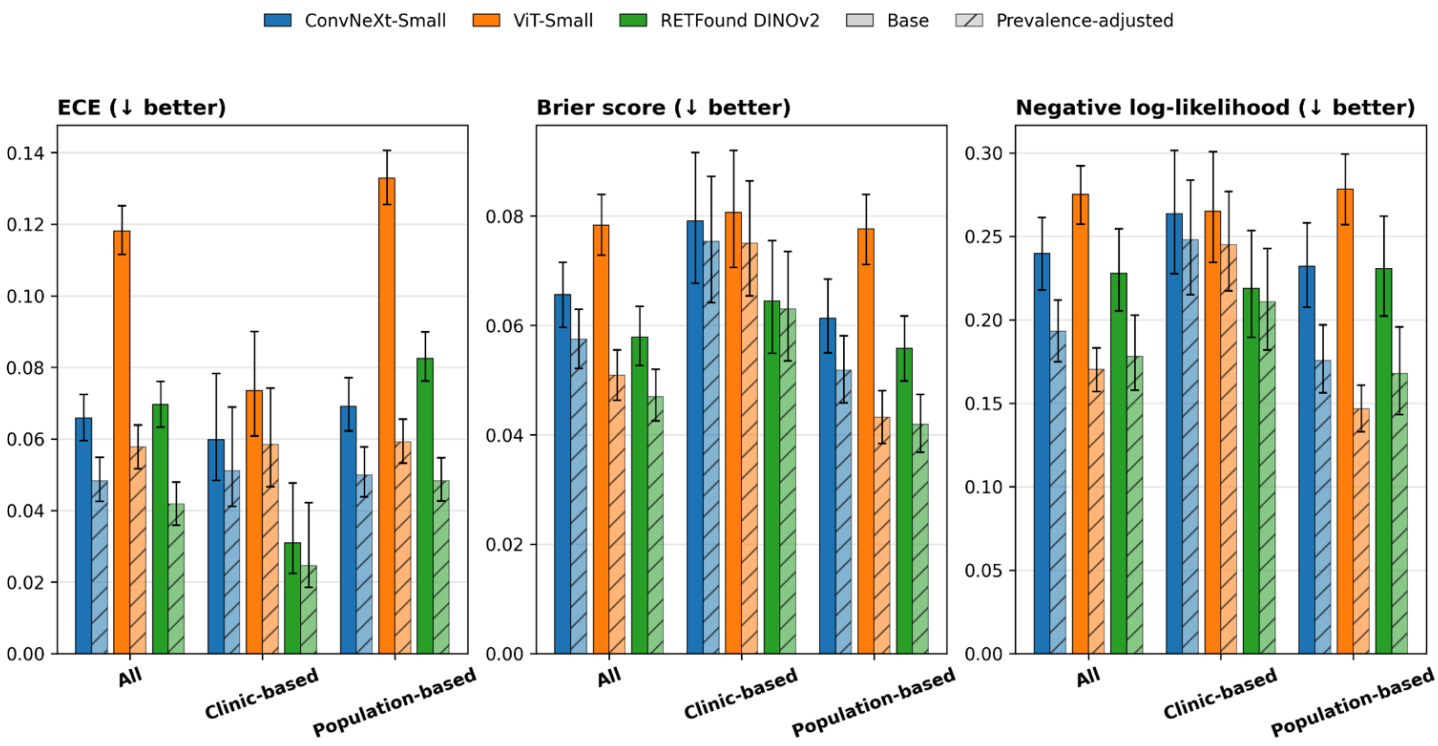


Figure 3

Probability calibration of three model architectures across three groupings of six external validation datasets, before and after prevalence-based odds adjustment.

The plot summarises the corresponding calibration metrics—ECE (left), Brier score (middle), and negative log-likelihood (right) with their 95% confidence intervals—comparing base predictions with prevalence-adjusted predictions across all three groups (all-datasets, clinic-based, and population-based) derived from six external validation datasets (IDRiD, HMADR_Clinic, UEMS, HMADR_Screening, CIEMS, and BES).

Abbreviations: BES, Beijing Eye Study; CIEMS, Central India Eye and Medical Study; ECE, expected calibration error; UEMS, Ural Eye and Medical Study.

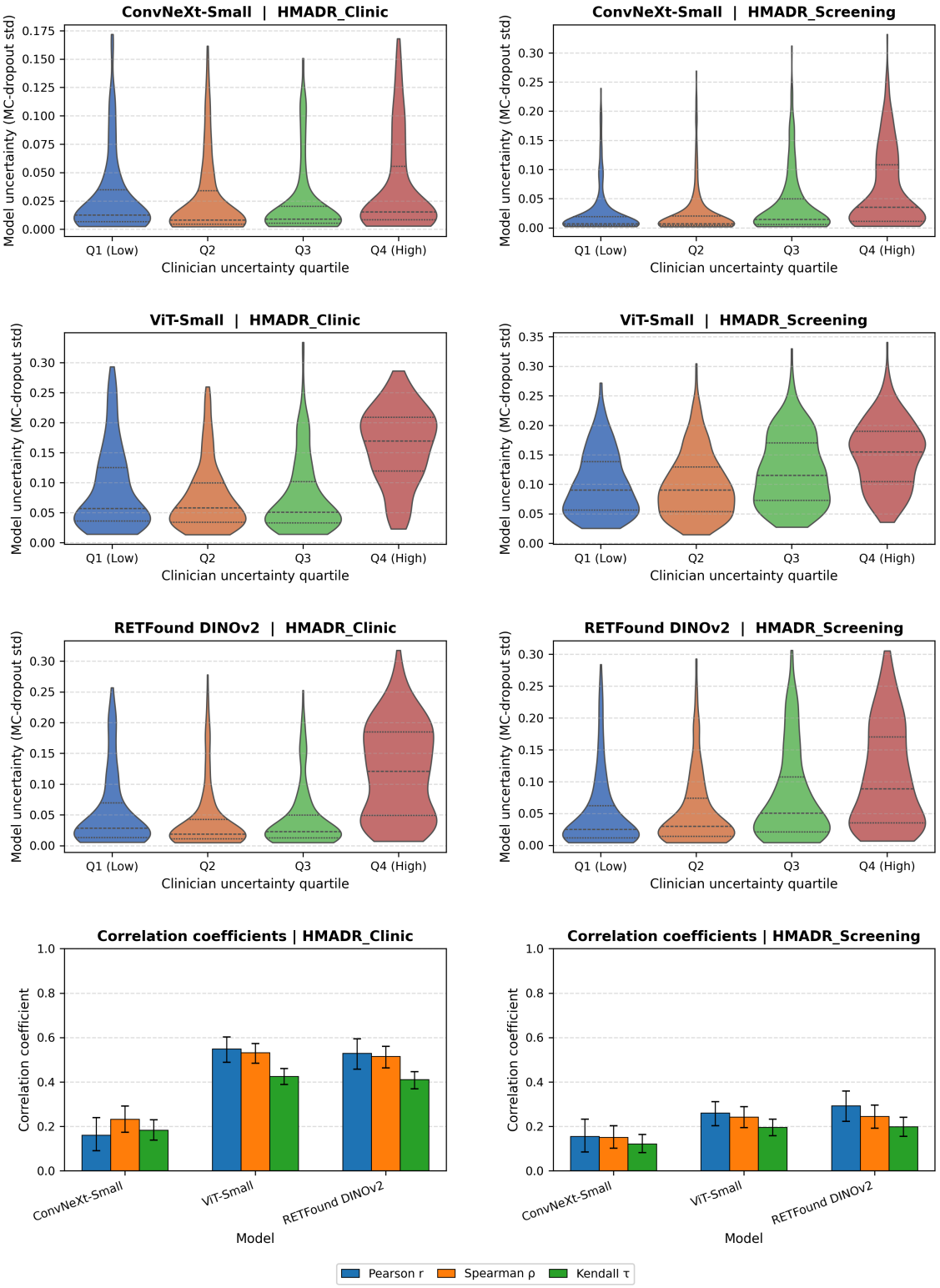


Figure 4

Alignment between model-derived and clinician-derived uncertainty across three model architectures and the HMADR datasets, with model uncertainty quantified by Monte Carlo dropout.

Each row in the upper panels corresponds to one model architecture (ConvNeXt-Small, ViT-Small, and RETFound DINOv2), and each pair of columns corresponds to one dataset (HMADR_Clinic and HMADR_Screening). The violin plots show the distribution of model uncertainty, defined as the predictive standard deviation estimated via Monte Carlo dropout, stratified by clinician uncertainty quartiles (Q1–Q4). Clinician uncertainty was defined as the standard deviation of the binary labels from five ophthalmologists for each image. The bottom row summarises the Pearson, Spearman's rank, and Kendall's tau correlation coefficients between clinician and model uncertainty for each model and dataset. Error bars indicate bootstrap 95% confidence intervals.

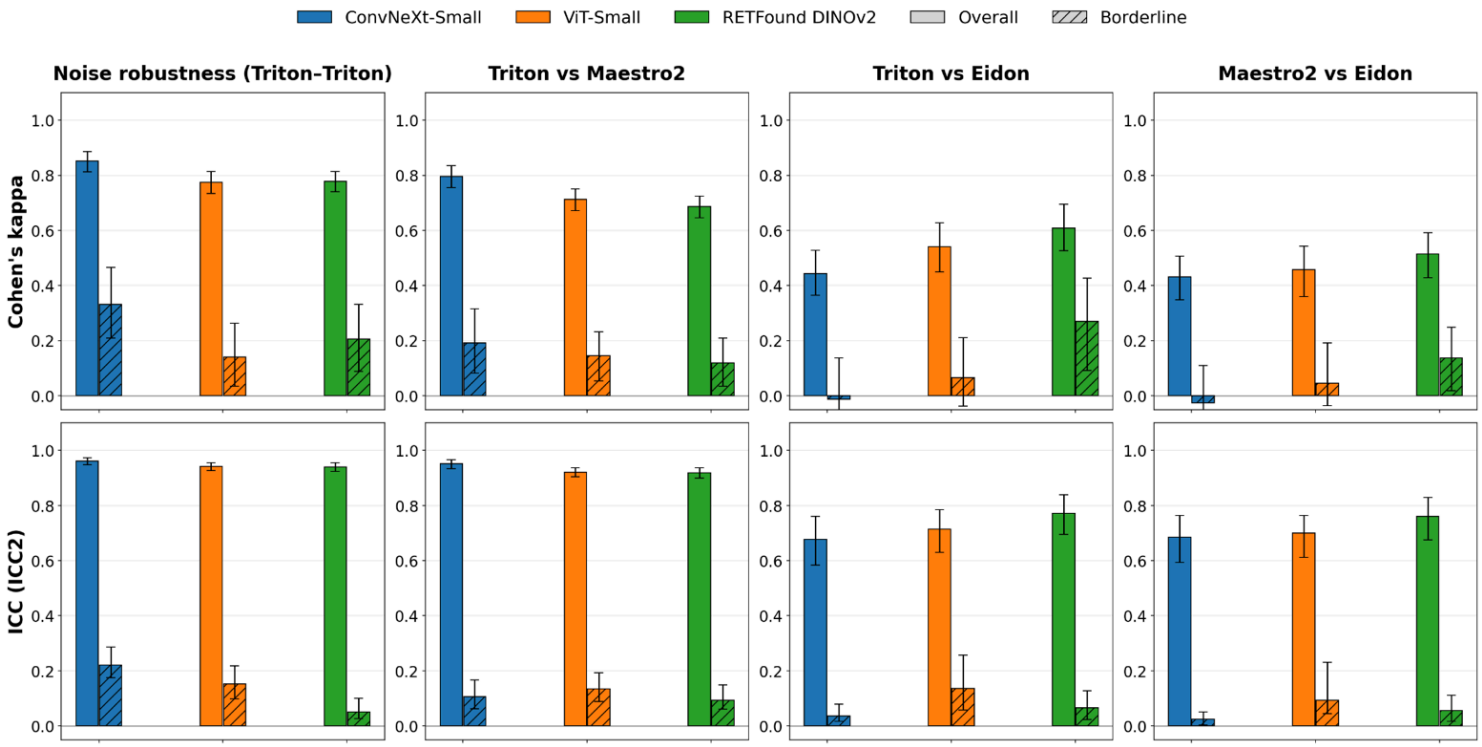


Figure 5

Robustness in Noise and Cross-Device Variation in the AI-READI Dataset

Bar charts summarise inter-prediction agreement for three model architectures (ConvNeXt-Small, ViT-Small, and RETFound DINOv2) across four comparison conditions: intra-device pairs for noise robustness (Triton paired acquisitions from the AI-READI dataset) and three cross-device pairs (Triton vs Maestro2, Triton vs Eidon, and Maestro2 vs Eidon). The upper row shows Cohen's kappa and the lower row shows ICC. Solid bars represent the overall paired dataset, whereas hatched bars represent the borderline subset, defined as cases within the upper or lower 10% of the decision threshold. Agreement was consistently higher in the overall dataset than in borderline cases across all conditions and model architectures.

Abbreviations: ICC, intraclass correlation coefficient.

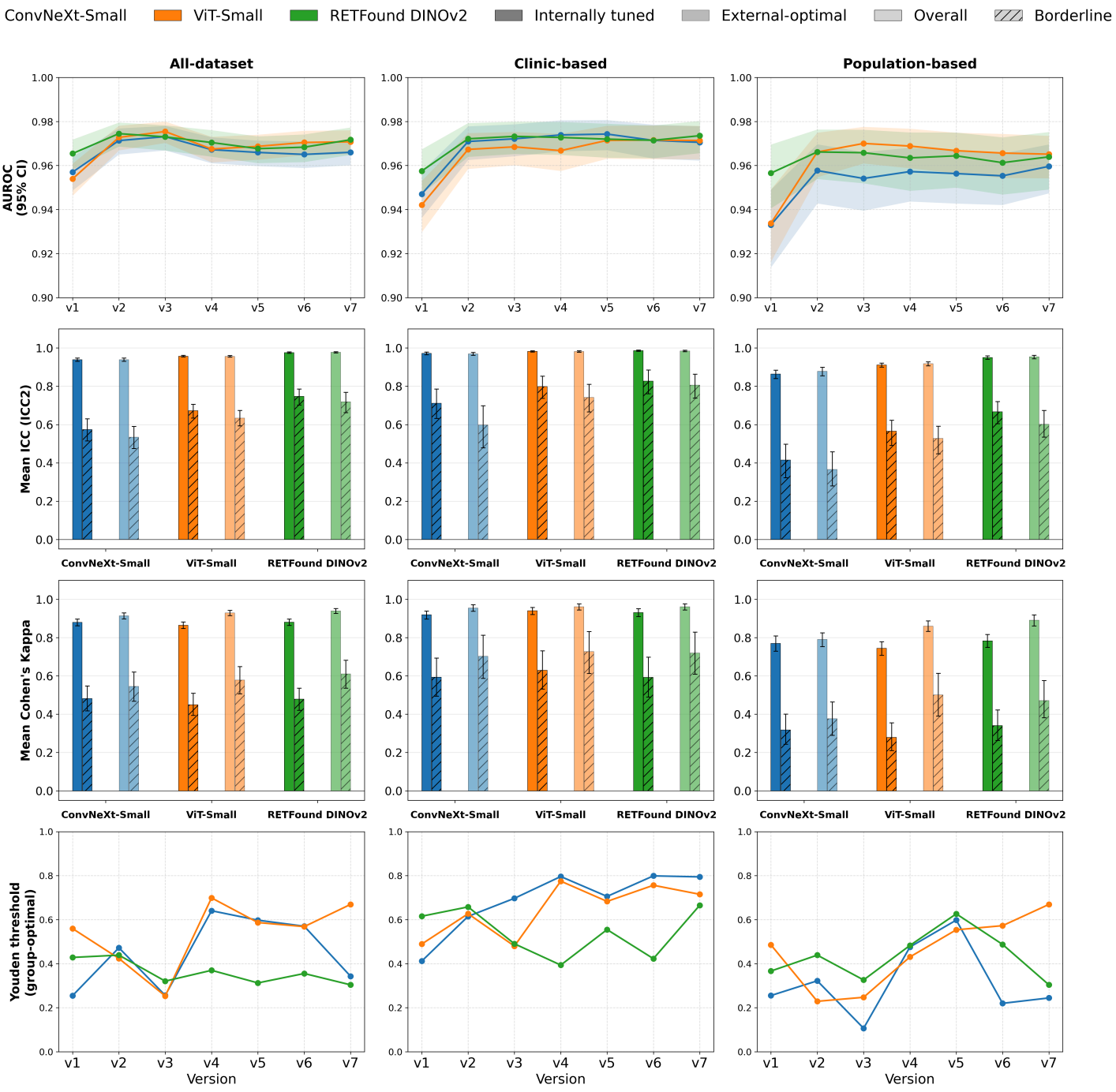


Figure 6

Inter-version agreement and decision-threshold stability across seven model updates of three model architectures, evaluated on three groupings of six external validation datasets.

Bar charts and line plots summarise discriminative performance (AUROC), inter-update agreement, and the stability of the Youden-optimal threshold across seven model versions (v1–v7) for three model architectures (ConvNeXt-Small, ViT-Small, and RETFound DINOv2), evaluated on three groupings (all-datasets, clinic-based, and population-based) derived from six external validation datasets (IDRiD, HMADR_Clinic, UEMS, HMADR_Screening, CIEMS, and BES). The top row shows AUROC across model versions with 95% bootstrap confidence intervals, plotted as separate lines for each architecture. The second and third rows show the average ICC and Cohen's kappa, respectively, between consecutive model versions (v1→v2 through v6→v7); solid bars represent the overall dataset, while hatched bars represent the borderline subset (cases within the upper or lower 10% of the decision threshold). The bottom row shows the group-specific Youden-optimal threshold plotted across model versions, with separate lines for each architecture.

Abbreviations: AUROC, area under the receiver operating characteristic curve; BES, Beijing Eye Study; CIEMS, Central India Eye and Medical Study; ICC, intraclass correlation coefficient; UEMS, Ural Eye and Medical Study.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- manuscript20260528final2supplementary.docx