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.

One-Year Prediction of Vision-Threatening Diabetic Retinopathy Complications Using Multimodal Artificial Intelligence: A Multicenter Real-World Study

Mostafa EL HABIB DAHO
Université de Bretagne Occidentale
Philippe ZHANG
Université de Bretagne Occidentale
Alireza REZAEI
Université de Bretagne Occidentale
Mohsen TABEJAMAAT
Université de Bretagne Occidentale
Mathieu LAMARD
Université de Bretagne Occidentale
Pascale MASSIN
Hôpital Lariboisière
Abir ZUREIK
Hôpital Lariboisière
Capucine LEPICARD
Hôpital Lariboisière
Eric VICAUT
Université Paris Cité
Mathias GALLARDO
Hôpital Lariboisière
Hugo LE BOITÉ
Université Paris Cité
Marine BONY
ADCIS S.A.
Pierre DEMAN
ADCIS S.A.
Aurélien DUCOURNAU
ADCIS S.A.
François POTEVIN
ADCIS S.A.

Laurent BORDERIE
Evolucare Technologies
Alexandre LE GUILCHER
Evolucare Technologies
Sejal GHATE
Carl Zeiss Meditec, Inc.
Ali SALEHI
Carl Zeiss Meditec, Inc.
Deborah COSETTE
Carl Zeiss Meditec AG
EviRed Investigators
Aude COUTURIER
Université Paris Cité
Gwenolé QUELLEC
gwenole.quellec@inserm.fr

Laboratoire de Traitement de l'Information Médicale

Article

Keywords:

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11433556/v1

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

Additional Declarations: Competing interest reported. A.C. serves as a consultant to ZEISS. G.Q. serves as a consultant to Evolucare Technologies and ADCIS. P.Z. has been employed by Evolucare Technologies. P.D., M.B., A.D. and F.P. are employees of ADCIS. L.B. and A.L.G. are employees of Evolucare Technologies. S.G. and A.S. are employees of Carl Zeiss Meditec, Inc. D.C. is an employee of Carl Zeiss Meditec AG. The remaining authors declare no competing interests.

One-Year Prediction of Vision-Threatening Diabetic Retinopathy Complications Using Multimodal Artificial Intelligence: A Multicenter Real-World Study

Mostafa El Habib Daho[1,2†], Philippe Zhang[1,2†], Alireza Rezaei[1,2], Mohsen Tabejamaat[1,2], Mathieu Lamard[1,2], Pascale Massin[3], Abir Zureik[3], Capucine Lepicard[3], Eric Vicaut[4,5], Mathias Gallardo[3], Hugo Le Boité[4,3], Marine Bony[6,7], Pierre Deman[6,7], Aurélien Ducournau[6,7], François Potevin[6,7], Laurent Borderie[7], Alexandre Le Guilcher[7], Sejal Ghate[8], Ali Salehi[8], Deborah Cosette[9], EviRed Investigators[§], Aude Couturier[‡4,3], Gwenolé Quellec[‡2*]

[1]Université de Bretagne Occidentale, 3 rue Matthieu Gallou, Brest, 29200, France.
[2]LaTIM UMR 1101, Inserm, 22 rue Camille Desmoulins, Brest, 29200, France.
[3]Département d'ophtalmologie, Hôpital Lariboisière, AP-HP, 2 rue Ambroise Paré, Paris, 75010, France.
[4]Université Paris Cité, 85 boulevard Saint-Germain, Paris, 75006, France.
[5]Unité de Recherche Clinique, Hôpital Lariboisière, AP-HP, 2 rue Ambroise Paré, Paris, 75010, France.
[6]ADCIS S.A., Saint-Contest, 14280, France.
[7]Evolucare Technologies, Villers-Bretonneux, 80800, France.
[8]Carl Zeiss Meditec, Inc., Dublin, CA 94568, USA.
[9]Carl Zeiss Meditec AG, Jena, 07743, Germany.

*Corresponding author(s). E-mail(s): gwenole.quellec@inserm.fr;

§ A full list of EviRed Investigators is provided in the Supplementary Information.
‡ These authors jointly supervised this work.

†These authors contributed equally to this work.

Abstract

Current diabetic retinopathy management relies primarily on contemporaneous severity grading and does not directly quantify the short-term risk of vision-threatening complications. We developed and internally evaluated image-based artificial intelligence models for one-year prediction of diabetic retinopathy severity, proliferative diabetic retinopathy (PDR) and center-involved diabetic macular edema (ci-DME) using retinal imaging acquired during routine care.

The prospective multicenter EviRed cohort included 3,183 patients with diabetes, 8,520 visits and 16,509 eye-level observations collected across 14 French hospitals. Device-specific ensembles were developed for ultra-widefield color fundus photography (CLARUS500 and OPTOS) and OCT angiography (CIRRUS5000 and PLEX Elite9000). Predictions from individual images and both eyes were aggregated at the patient-visit level. Participants were partitioned at the patient level into a development set (65%) and a held-out test set (35%).

For prediction of diabetic retinopathy severity at the subsequent annual visit, OPTOS and CLARUS500 achieved quadratic weighted kappa values of 0.866 (95% confidence interval, 0.842–0.878) and 0.852 (0.830–0.877), respectively. For incident ci-DME, CIRRUS5000 achieved the highest discrimination (AUC, 0.826; 0.770–0.878), whereas PLEX Elite9000 yielded the highest AUC point estimate for incident PDR (0.877; 0.807–0.950). Late fusion did not consistently outperform the strongest device-specific model. In secondary complete-case analyses, observed ci-DME incidence increased across quartiles of the image-derived risk score, reaching 19.7–30.4% in the highest-risk quartile. AI models also achieved higher AUC point estimates than unassisted expert risk assessments for both complications, although these comparisons were exploratory.

Retinal imaging acquired at a single routine-care visit therefore contains information associated with diabetic retinopathy severity and vision-threatening complications one year later. These findings support image-based short-term risk stratification, while external validation, calibration and prospective clinical utility assessment remain necessary before deployment.

Keywords: Diabetic retinopathy, Proliferative diabetic retinopathy, Diabetic macular edema, Multimodal learning, Longitudinal modeling, Risk prediction

1 Introduction

Diabetic retinopathy (DR) remains a leading cause of preventable blindness worldwide and represents a major public health burden despite the availability of effective treatments. Recent global estimates indicate that tens of millions of individuals are affected by DR and its vision-threatening complications, including proliferative diabetic retinopathy (PDR) and diabetic macular edema [1, 2]. As a common microvascular complication of diabetes, DR remains a major cause of vision loss among working-age adults [3].

Current clinical management relies primarily on severity grading systems, such as the International Clinical Diabetic Retinopathy (ICDR) scale, which categorize disease according to retinal findings observed at a given time point [4]. These systems have been instrumental in standardizing diagnosis, communication and treatment decisions. However, they provide a static description of a dynamic disease and do not directly quantify an individual's short-term risk of progression.

This limitation is clinically important because patients with similar baseline grades may have substantially different risks of developing vision-threatening complications, including PDR and center-involved diabetic macular edema (ci-DME). Consequently, follow-up intervals remain largely determined by broad severity categories rather than individualized risk estimates [5], potentially resulting in delayed identification of high-risk patients or unnecessary monitoring of patients at lower risk.

Advances in retinal imaging and automated image analysis have enabled increasingly detailed characterization of retinal lesions, vascular abnormalities and structural changes [6]. Different imaging modalities capture complementary manifestations of diabetic retinal disease. Ultra-widefield color fundus photography (UWF-CFP) provides extensive visualization of retinal lesions and peripheral abnormalities, whereas optical coherence tomography (OCT) and OCT angiography (OCT-A) characterize macular structure and retinal microvasculature [7]. Nevertheless, most artificial intelligence systems for DR remain focused on cross-sectional detection or severity grading rather than future clinical outcomes.

Recent studies have demonstrated the potential of machine learning to predict DR progression and vision-threatening complications from retinal images and clinical variables [8–11]. These findings support a transition from static disease classification towards individualized risk stratification. However, systematic reviews indicate that many existing models have been developed using relatively small, curated or single-modality datasets, with limited evaluation under heterogeneous real-world conditions [12–14].

Routine clinical environments introduce several additional challenges. Data are acquired across multiple centers and imaging devices, acquisition protocols vary, follow-up intervals may be irregular, and not all imaging modalities are available at every visit. Clinical annotations may also differ between local assessments and centralized reading. Previous multicenter studies have shown that AI performance can vary across centers, devices and acquisition conditions, emphasizing the importance of evaluation in representative clinical populations [15, 16]. A prediction system intended for routine care should therefore accommodate device-specific information and remain operational when only a subset of imaging modalities is available.

In this context, longitudinal prediction refers to the use of information available at a current visit to estimate disease status or complication risk at a subsequent time point. Rather than explicitly modeling a sequence of previous visits, this setting addresses a clinically actionable question: whether retinal images acquired today contain information that can identify patients at increased risk of developing vision-threatening disease within the following year. This fixed-horizon prognostic approach may provide information that is not captured by the current severity grade alone.

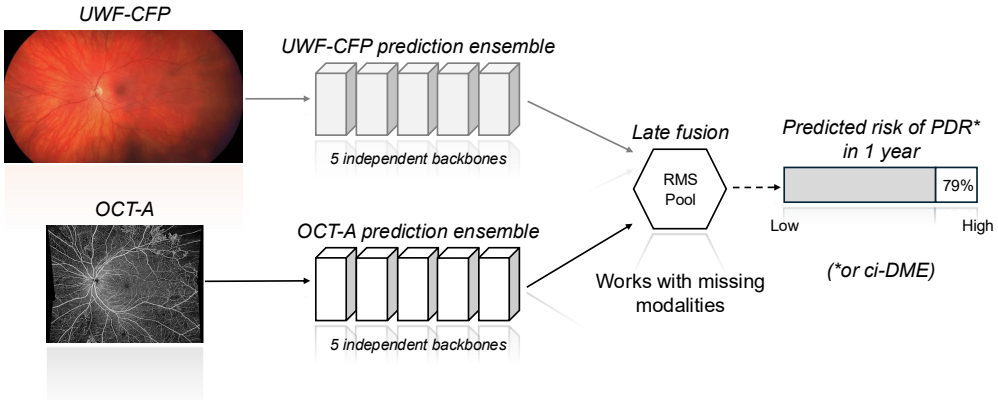


Fig. 1: Overview of the multimodal longitudinal prediction framework. Multimodal data are collected at each visit within a multicenter clinical setting, including ultra-widefield color fundus photography (UWF-CFP) and optical coherence tomography angiography (OCT-A). The model is designed to handle missing modalities and irregular follow-up intervals, reflecting real-world clinical conditions. All predictions were performed at the patient-visit level, including the risk of vision-threatening diabetic retinopathy (DR) complications within one year.

In this study, we used prospectively collected data from the multicenter EviRed cohort to develop and internally evaluate image-based models for three complementary tasks: cross-sectional DR severity assessment, prediction of DR severity at the subsequent annual visit, and one-year prediction of PDR and ci-DME. Models processed the retinal images available at each index visit, and image-level predictions were aggregated across acquisitions and both eyes to obtain a patient-visit-level estimate using a worst-eye strategy. Separate ensembles were developed for CLARUS500, OPTOS, CIRRUS5000 and PLEX Elite9000, while late fusion combined available device-level predictions without requiring complete multimodal acquisition.

Clinical variables were not used as inputs to the prediction models. Instead, they were retained to contextualize the image-derived risk estimates within clinician-facing reports and to support human interpretation of the predicted risk. We investigated whether retinal imaging acquired during routine care could support individualized one-year risk stratification, while remaining applicable across heterogeneous devices, centers and patterns of modality availability (Fig. 1).

2 Results

2.1 Study population and real-world data availability

A total of 3,183 patients with diabetes were included, contributing 8,520 visits and 16,509 eye-level observations. The median follow-up duration was 2.0 years (interquartile range, 1.0–3.0), with an average of 2.7 visits per patient.

Baseline demographic, systemic and ophthalmic characteristics of the cohort, together with the distributions of recruiting centres and imaging examinations recorded at inclusion, are summarised in Table 1.

Retinal imaging collected during routine care included ultra-widefield color fundus photography (UWF-CFP) and optical coherence tomography and angiography (OCT-A), acquired using multiple devices across participating centers. Systemic clinical variables were collected concurrently for cohort characterization and contextual reporting but were not used as inputs to the prediction models.

Although imaging data were initially processed at the image and eye levels, predictions and reference labels were aggregated across both eyes using a worst-eye strategy. Model performance was therefore evaluated at the patient-visit level, with each eligible visit representing a distinct prediction time point.

Imaging availability varied across visits because of incomplete acquisitions and irregular follow-up. UWF-CFP was unavailable for 23.2% of observations, whereas OCT-A was unavailable for 32.5%, reflecting the heterogeneous acquisition patterns encountered in routine clinical care.

Reference annotations were obtained from local investigators and centralized reading. Exact agreement between local and centralized DR grades was 55.4%, illustrating substantial variability in real-world clinical labeling.

The cohort was partitioned at the patient level into a development set (65%) and a held-out test set (35%). All images, eyes, and repeated visits from a given patient were assigned to the same partition, preventing patient-level overlap between the two sets. The test set remained strictly isolated throughout model development and selection.

Table 1: Baseline characteristics of the EviRed cohort.

Characteristic	Overall (N = 3,183)	Unavailable, n (%)
Demographic and systemic characteristics		
Patients, n	3,183	0 (0.0)
Age, years	58.9 ± 13.9	0 (0.0)
Sex, n (%)		0 (0.0)
Male	1,988 (62.5)	
Female	1,195 (37.5)	
Diabetes type, n (%)		27 (0.8)
Type 1	766 (24.1)	
Type 2	2,349 (73.8)	
Other	41 (1.3)	
HbA1c, %	8.2 ± 1.8	558 (17.5)
Body mass index, kg/m^2	28.6 ± 5.8	325 (10.2)
Ophthalmic characteristics at baseline		
Local DR status, worst available eye, n (%)		28 (0.9)
No DR	494 (15.5)	
Mild NPDR	290 (9.1)	
Moderate NPDR	1,136 (35.7)	
Severe NPDR	565 (17.8)	

Continued on next page

Characteristic	Overall (N = 3,183)	Unavailable, n (%)
PDR	670 (21.0)	
Imaging examinations recorded at M0, n (%)		
CLARUS500 (UWF-CFP)	2,920 (91.7)	28 (0.9)
OPTOS (UWF-CFP)	2,346 (73.7)	30 (0.9)
PLEX Elite9000 (OCT/OCT-A)	2,890 (90.8)	29 (0.9)
CIRRUS5000 (OCT/OCT-A)	1,790 (56.2)	38 (1.2)
Recruiting centre, n (%)		
001 — Paris-APHP - Lariboisière	505 (15.9)	
002 — Dijon-CHRU de Dijon - Bocage	346 (10.9)	
003 — Bordeaux-CHU de Bordeaux - Pellegrin	139 (4.4)	
004 — Lyon-HC de Lyon - La Croix Rousse	264 (8.3)	
005 — Bobigny-APHP - Avicenne	310 (9.7)	
006 — Créteil-CHI Créteil	157 (4.9)	
007 — Paris-APHP - Pitié Salpêtrière	293 (9.2)	
008 — Paris-Centre de Santé Broca	502 (15.8)	
009 — Brest-CHU de Brest - Morvan	177 (5.6)	
010 — Paris-CHNO des Quinze-Vingts	260 (8.2)	
011 — Nantes-CHU de Nantes - Hôtel Dieu	38 (1.2)	
012 — Paris-Fondation Ophtalmologique A. de Rothschild	101 (3.2)	
013 — Marseille-Centre Monticelli Paradis d'Ophtalmologie	76 (2.4)	
014 — Nice-CHU de Nice - Pasteur 2	15 (0.5)	

2.2 Prediction tasks and evaluation setting

We defined three complementary prediction tasks, progressing from the assessment of current disease severity to the prediction of clinically actionable outcomes one year later (Fig. 2).

First, we evaluated cross-sectional classification of DR severity at the index visit t. This task provided a reference assessment of the models' ability to reproduce the grading information used in routine clinical practice and enabled comparison with conventional automated DR grading approaches.

Second, we evaluated prediction of DR severity at the subsequent annual visit $t+1$ using only retinal images available at the index visit. This task assessed whether current imaging contained information associated with the patient's future disease state approximately one year later.

Third, we evaluated the one-year prediction of two vision-threatening complications, PDR and ci-DME. These outcomes were analyzed separately and constituted the primary clinically actionable prediction tasks of the study.

For each complication-specific analysis, only patient-visits at risk of developing the corresponding outcome were included. ci-DME prediction was restricted to index visits without ci-DME in either eye, whereas PDR prediction was restricted to index visits without PDR in either eye.

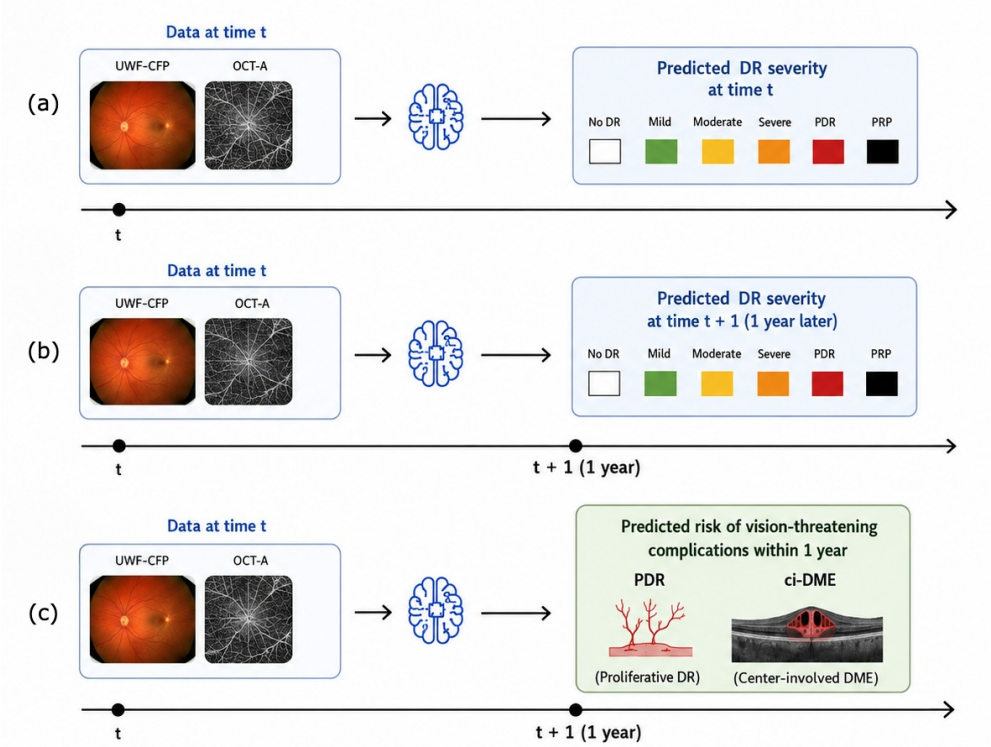


Fig. 2: Definition of the prediction tasks. Three complementary settings were evaluated using retinal images available at an index visit t. **a**, Cross-sectional classification of diabetic retinopathy (DR) severity at the index visit. **b**, Prediction of DR severity at the subsequent annual visit $t+1$. **c**, Prediction of vision-threatening complications within one year, including proliferative diabetic retinopathy (PDR) and center-involved diabetic macular edema (ci-DME). PDR and ci-DME prediction constituted the primary clinically actionable tasks of the study.

Predictions were initially generated from individual images and subsequently aggregated within each visit. Information from the right and left eyes was combined using a worst-eye strategy, resulting in one prediction and one reference label for each eligible patient-visit. For ordinal severity tasks, the reference label corresponded to the highest DR severity observed in either eye. For binary outcomes, a patient-visit was considered positive if the complication occurred in either eye at the subsequent annual visit.

A patient could contribute multiple eligible visit-to-visit transitions, provided that sufficient follow-up was available to determine the outcome at $t + 1$. However, all images, eyes and repeated visits from the same patient were assigned to the same data partition.

Performance was evaluated once on the held-out test set, with no patient overlap between the development and test sets. The area under the receiver operating characteristic curve (AUC) was used for PDR and ci-DME prediction, whereas quadratic weighted kappa (QWK) and accuracy were used for ordinal DR severity assessment. All metrics were calculated at the patient-visit level.

Device-specific models and late-fusion configurations were compared to characterize the relative contribution of the available imaging modalities and the ability of the framework to operate across different patterns of modality availability. Model performance was also compared with unassisted clinical expert risk estimates for PDR and ci-DME prediction.

The data partitions, outcome definitions and evaluation metrics were specified before final model evaluation, and the held-out test set was not used for model selection or hyperparameter optimization.

2.3 Cross-sectional classification of DR severity

We first evaluated the ability of the device-specific and late-fusion models to classify DR severity at the index visit. This cross-sectional task served as a reference analysis to determine whether the models captured the imaging features represented in the operational EviRed grading scheme before evaluating future disease states.

Performance varied across imaging modalities (Table 2). The UWF-CFP models achieved the highest agreement with the reference grades. CLARUS500 yielded a QWK of 0.889 (95% CI, 0.881–0.898) and an accuracy of 0.656 (95% CI, 0.640–0.681), whereas OPTOS achieved a QWK of 0.891 (95% CI, 0.871–0.905) and an accuracy of 0.628 (95% CI, 0.602–0.653).

Performance was lower for the OCT-A models. PLEX Elite9000 achieved a QWK of 0.782 (95% CI, 0.750–0.807) and an accuracy of 0.505 (95% CI, 0.478–0.527), whereas CIRRUS5000 achieved a QWK of 0.641 (95% CI, 0.579–0.675) and an accuracy of 0.330 (95% CI, 0.293–0.356).

The late-fusion configurations showed intermediate performance. CLARUS500–CIRRUS5000 achieved a QWK of 0.825 (95% CI, 0.806–0.847), whereas CLARUS500–PLEX Elite9000 achieved a QWK of 0.863 (95% CI, 0.834–0.875). Neither configuration exceeded the strongest UWF-CFP model, indicating that UWF-CFP provided the imaging information most directly aligned with cross-sectional DR severity grading.

The row-normalized confusion matrices showed that most UWF-CFP predictions remained close to the reference grade (Fig. 3). For CLARUS500 and OPTOS, misclassifications occurred predominantly between adjacent or clinically related categories. By contrast, OCT-A predictions were more widely distributed across intermediate and advanced grades. These patterns are consistent with the closer correspondence between color fundus findings and the retinal lesions used to define conventional DR severity categories.

2.4 One-year prediction of future DR severity

We next evaluated whether retinal images acquired at the index visit could predict DR severity at the subsequent annual visit. Predictive performance was generally lower

Table 2: **Performance for cross-sectional classification and one-year prediction of diabetic retinopathy severity.** Values are point estimates with patient-clustered 95% confidence intervals. QWK, quadratic weighted kappa.

Imaging configuration	Index visit classification		One-year prediction	
	QWK (95% CI)	Accuracy (95% CI)	QWK (95% CI)	Accuracy (95% CI)
CLARUS500	0.889 (0.881–0.898)	0.656 (0.640–0.681)	0.852 (0.830–0.877)	0.584 (0.558–0.622)
OPTOS	0.891 (0.871–0.905)	0.628 (0.602–0.653)	0.866 (0.842–0.878)	0.601 (0.566–0.634)
CIRRUS5000	0.641 (0.579–0.675)	0.330 (0.293–0.356)	0.648 (0.586–0.693)	0.304 (0.266–0.346)
PLEX Elite9000	0.782 (0.750–0.807)	0.505 (0.478–0.527)	0.739 (0.717–0.769)	0.477 (0.453–0.498)
CLARUS500 + Cirrus	0.825 (0.806–0.847)	0.542 (0.521–0.558)	0.799 (0.769–0.823)	0.497 (0.472–0.525)
CLARUS500 + PLEX Elite9000	0.863 (0.834–0.875)	0.589 (0.561–0.601)	0.835 (0.810–0.863)	0.553 (0.522–0.589)

than for cross-sectional classification, although the magnitude of the difference varied across imaging configurations (Table 2).

UWF-CFP yielded the highest performance for future severity prediction. OPTOS achieved a QWK of 0.866 (95% CI, 0.842–0.878) and an accuracy of 0.601 (95% CI, 0.566–0.634), whereas CLARUS500 achieved a QWK of 0.852 (95% CI, 0.830–0.877) and an accuracy of 0.584 (95% CI, 0.558–0.622). Relative to cross-sectional classification, QWK decreased by 0.025 for OPTOS and by 0.037 for CLARUS500, indicating that both UWF-CFP models retained substantial agreement with the future reference grade.

Performance was lower for OCT-A. PLEX Elite9000 achieved a QWK of 0.739 (95% CI, 0.717–0.769) and an accuracy of 0.477 (95% CI, 0.453–0.498), compared with a cross-sectional QWK of 0.782. CIRRUS5000 achieved a QWK of 0.648 (95% CI, 0.586–0.693) and an accuracy of 0.304 (95% CI, 0.266–0.346). Its QWK was similar to that observed for cross-sectional classification, although exact classification accuracy remained lower.

The late-fusion configurations achieved intermediate performance. CLARUS500–CIRRUS5000 yielded a QWK of 0.799 (95% CI, 0.769–0.823) and an accuracy of 0.497 (95% CI, 0.472–0.525), whereas CLARUS500–PLEX Elite9000 achieved a QWK of 0.835 (95% CI, 0.810–0.863) and an accuracy of 0.553 (95% CI, 0.522–0.589). As in the cross-sectional task, neither fusion configuration exceeded the strongest UWF-CFP model.

The row-normalized confusion matrices showed that predictions from the UWF-CFP models were predominantly distributed around the future reference grade (Fig. 4). Most errors involved adjacent or clinically related categories, whereas OCT-A predictions were more widely distributed across intermediate and advanced grades. Misclassification was more pronounced for the less frequent advanced categories, for which predictions were frequently assigned to neighboring severity levels.

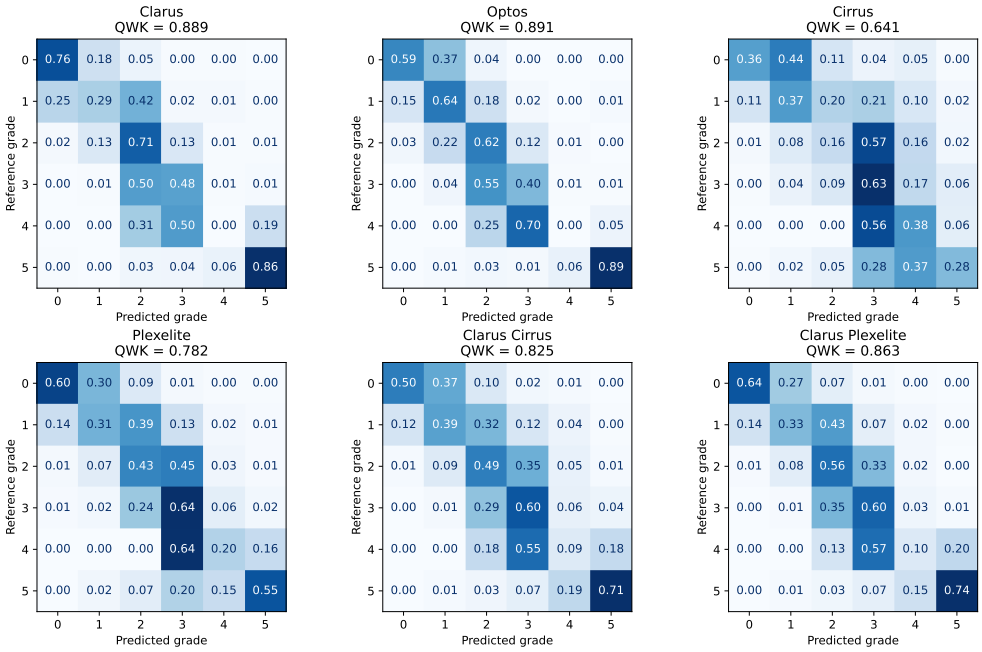


Fig. 3: **Cross-sectional classification of diabetic retinopathy severity at the index visit.** Row-normalized confusion matrices are shown for the four device-specific models and two late-fusion configurations. Rows indicate reference grades and columns indicate predicted grades according to the six-category operational EviRed grading scheme. Values represent the proportion of patient-visits within each reference category. Quadratic weighted kappa values are reported above each matrix.

Together, these results show that retinal images acquired at a single visit contain information associated with DR severity at the subsequent annual visit. The limited reduction in QWK for the UWF-CFP models supports the feasibility of short-term prognostic assessment beyond contemporaneous disease grading.

2.5 One-year prediction of incident vision-threatening complications

We next evaluated whether retinal images acquired at the index visit could predict the development of vision-threatening DR complications at the subsequent annual visit. Two clinically actionable outcomes were evaluated separately: incident center-involved diabetic macular edema (ci-DME) and incident proliferative diabetic retinopathy (PDR).

To focus on the prediction of new complications, each analysis was restricted to patient-visits at risk for the corresponding outcome. ci-DME prediction included only patient-visits without ci-DME in either eye at the index visit, whereas PDR prediction included only patient-visits without PDR in either eye at the index visit. Model

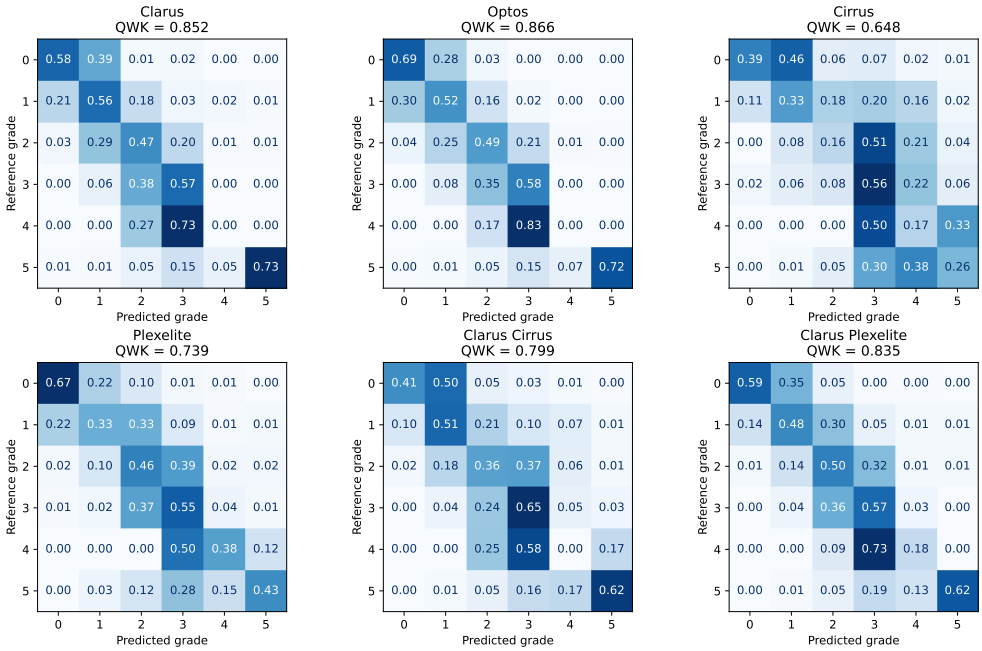


Fig. 4: **One-year prediction of diabetic retinopathy severity.** Row-normalized confusion matrices for the four device-specific models and two late-fusion configurations. Retinal images acquired at the index visit were used to predict DR severity at the subsequent annual visit. Rows indicate future reference grades and columns indicate predicted grades according to the six-category operational EviRed grading scheme. Values represent the proportion of patient-visits within each reference category. Quadratic weighted kappa values are reported above each matrix.

performance was evaluated at the patient-visit level using the area under the receiver operating characteristic curve (AUC).

Discrimination varied according to both the predicted outcome and the available imaging configuration (Table 3; Fig. 5).

For ci-DME, CIRRUS5000 achieved the highest discrimination, with an AUC of 0.826 (95% CI, 0.770–0.878), followed by PLEX Elite9000 with an AUC of 0.789 (95% CI, 0.736–0.832). The UWF-CFP models yielded AUCs of 0.753 (95% CI, 0.688–0.774) for CLARUS500 and 0.687 (95% CI, 0.637–0.736) for OPTOS. The CLARUS500–CIRRUS5000 and CLARUS500–PLEX Elite9000 late-fusion configurations achieved AUCs of 0.750 (95% CI, 0.721–0.799) and 0.788 (95% CI, 0.763–0.818), respectively. The higher point estimates obtained with CIRRUS5000 and PLEX Elite9000 suggest that OCT-A-derived information was particularly informative for identifying patient-visits at increased risk of developing ci-DME within one year.

For PDR, PLEX Elite9000 yielded the highest AUC point estimate at 0.877 (95% CI, 0.807–0.950). The OPTOS–PLEX Elite9000 late-fusion configuration achieved an AUC of 0.827 (95% CI, 0.740–0.934), followed by CLARUS500 at 0.826 (95% CI,

Table 3: One-year prediction of incident vision-threatening diabetic retinopathy complications. Values are areas under the receiver operating characteristic curve with 95% confidence intervals. AUC, area under the receiver operating characteristic curve; ci-DME, center-involved diabetic macular edema; PDR, proliferative diabetic retinopathy.

Imaging configuration	ci-DME AUC (95% CI)	PDR AUC (95% CI)
CLARUS500	0.753 (0.688–0.774)	0.826 (0.756–0.906)
OPTOS	0.687 (0.637–0.736)	0.779 (0.676–0.900)
CIRRUS5000	0.826 (0.770–0.878)	0.744 (0.565–0.908)
PLEX Elite9000	0.789 (0.736–0.832)	0.877 (0.807–0.950)
CLARUS500 + CIRRUS5000	0.750 (0.721–0.799)	0.692 (0.565–0.813)
OPTOS + PLEX Elite9000	0.775 (0.726–0.823)	0.827 (0.740–0.934)
CLARUS500 + PLEX Elite9000	0.788 (0.763–0.818)	0.807 (0.734–0.900)

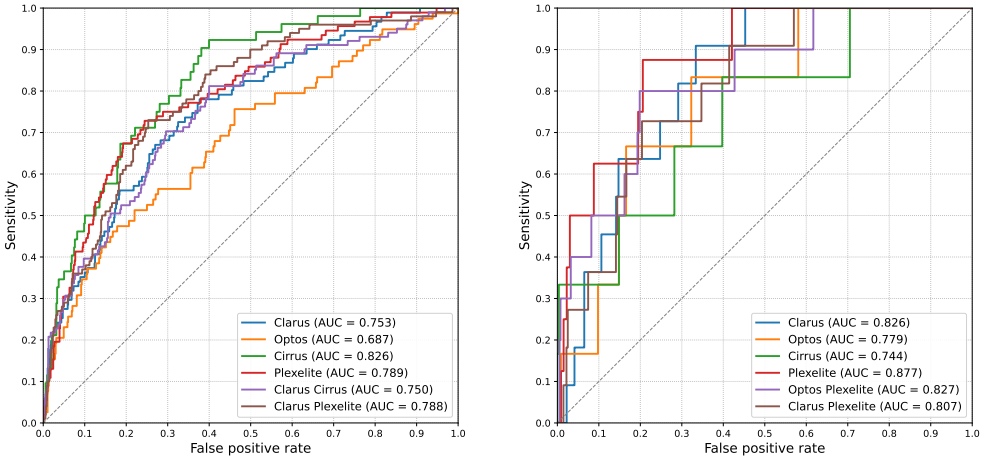
0.756–0.906), CLARUS500–PLEX Elite9000 at 0.807 (95% CI, 0.734–0.900), OPTOS at 0.779 (95% CI, 0.676–0.900) and CIRRUS5000 at 0.744 (95% CI, 0.565–0.908). Confidence intervals were wider for PDR than for ci-DME, reflecting the lower number of incident PDR events and the resulting uncertainty in the performance estimates. The overlapping confidence intervals do not support a formal conclusion that one imaging configuration was statistically superior to the others.

Late fusion did not systematically improve discrimination over the strongest device-specific model. Its principal contribution was therefore operational rather than a uniform increase in predictive performance. The framework provided a common mechanism for combining predictions when several imaging modalities were available, while retaining the ability to generate a device-specific prediction when only one modality had been acquired. This flexibility enabled risk estimation across the heterogeneous and incomplete imaging patterns encountered in the EviRed cohort.

Overall, retinal images acquired during routine care contained information associated with the development of vision-threatening complications at the subsequent annual visit. The imaging configuration providing the highest discrimination depended on the predicted endpoint, supporting an outcome-dependent and availability-aware approach to patient-visit-level risk stratification.

2.6 Outcome-dependent modality contributions and prediction with incomplete imaging

No single imaging modality provided the highest performance across all prediction tasks. UWF-CFP yielded the highest QWK values for both cross-sectional and one-year DR severity assessment, whereas OCT-A provided the highest discrimination for the vision-threatening outcomes. CIRRUS5000 achieved the highest AUC for ci-DME prediction, while PLEX Elite9000 yielded the highest point estimate for PDR. These



(a) Incident ci-DME.

(b) Incident PDR.

Fig. 5: One-year prediction of incident vision-threatening diabetic retinopathy complications. Receiver operating characteristic curves for the prediction of **a**, incident center-involved diabetic macular edema (ci-DME), and **b**, incident proliferative diabetic retinopathy (PDR), at the subsequent annual visit. Analyses were restricted to patient-visits without the corresponding complication in either eye at the index visit. Retinal images acquired at the index visit were used to generate patient-visit-level risk estimates. Curves are shown for the evaluated device-specific models and late-fusion configurations.

findings indicate that the relative contribution of each modality was outcome dependent, consistent with the distinct structural, vascular and lesion-level information captured by the different imaging technologies.

Late-fusion configurations yielded intermediate performance and did not systematically exceed the strongest device-specific model. Their principal contribution was therefore operational rather than a uniform improvement in predictive accuracy. When several modalities were available, their device-level predictions could be combined within a common framework; when only one modality was available, the corresponding device-specific prediction remained usable without requiring image imputation or complete multimodal acquisition.

This flexibility was particularly relevant in the EviRed cohort, in which UWF-CFP and OCT-A were unavailable for 23.2% and 32.5% of observations, respectively. The framework therefore extended prediction across heterogeneous acquisition patterns while preserving the outcome-specific information provided by each imaging modality.

Overall, these results support an availability-aware and outcome-specific use of multimodal retinal imaging. They also show that the clinical value of multimodality should not be defined solely by whether fusion surpasses the best unimodal model, but also by its ability to provide standardized predictions across variable and incomplete imaging configurations.

2.7 Comparison with baseline risk indicators and risk stratification

We performed a secondary complete-case analysis to compare image-based AI predictions with individual baseline risk indicators for one-year ci-DME prediction. The analysis was restricted to patient-visits with complete data for the AI-derived risk estimate, HbA1c, baseline DME status and baseline DR severity. Consequently, the evaluated subsets were smaller than those used in the primary analyses, and the AUC estimates reported here should not be directly compared with those obtained in the full eligible cohort.

Across all four imaging devices, AI-derived predictions yielded higher AUC point estimates than each individual baseline indicator (Fig. 6). For CLARUS500, the AI prediction achieved an AUC of 0.767, compared with 0.695 for baseline DR severity, 0.670 for baseline DME status and 0.652 for HbA1c. The corresponding AUCs for OPTOS were 0.737, 0.702, 0.673 and 0.657, respectively. Larger differences were observed for the OCT-A devices. For CIRRUS5000, the AI prediction achieved an AUC of 0.863, compared with 0.743 for baseline DR severity, 0.674 for baseline DME status and 0.647 for HbA1c. For PLEX Elite9000, the corresponding AUCs were 0.830, 0.721, 0.685 and 0.642.

Observed ci-DME incidence increased across quartiles of the AI-derived risk score for each device. In the lowest and highest risk quartiles, incidence was 3.1% and 24.7% for CLARUS500, 2.4% and 19.7% for OPTOS, 0.0% and 26.7% for Cirrus, and 1.3% and 30.4% for PLEX Elite9000, respectively. Intermediate quartiles generally showed progressively increasing event rates, supporting the ability of the model to separate patient-visits into groups with different observed one-year risks.

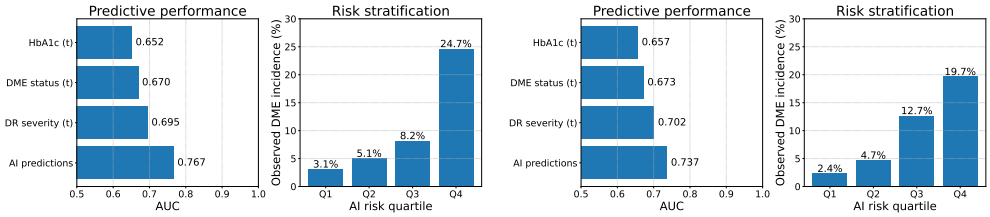
These analyses indicate that, within the complete-case subsets, the image-derived predictions provided greater discrimination than individual baseline clinical indicators and generated clinically interpretable risk strata. However, comparisons between separate AUC point estimates do not establish incremental predictive value beyond a combined multivariable clinical model. In addition, complete-case restriction may have introduced selection bias and limited direct comparison with the primary cohort analyses.

2.8 Comparison with clinical experts

We compared the one-year predictive performance of the AI models with unassisted risk estimates provided by clinical experts for incident PDR and ci-DME.

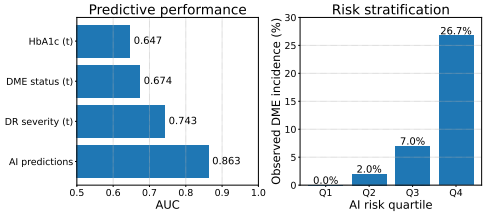
Expert assessments were obtained from retinal specialists at participating centers using the clinical and imaging information available at the index visit. Model-derived and expert-derived risk estimates were evaluated on the same held-out test cases using the area under the receiver operating characteristic curve (AUC).

Across all evaluated imaging configurations, the AI models achieved higher AUC point estimates than the corresponding expert assessments (Fig. 7). For PDR, model AUCs ranged from 0.744 to 0.877, compared with 0.544 to 0.578 for expert-derived risk estimates. For ci-DME, model AUCs ranged from 0.687 to 0.826, whereas expert AUCs ranged from 0.483 to 0.536.

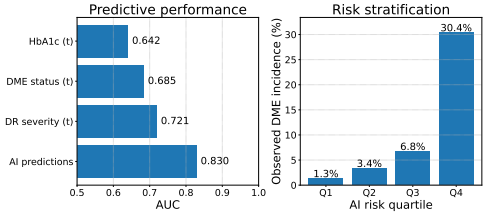


(a) CLARUS500.

(b) OPTOS.



(c) CIRRUS5000.

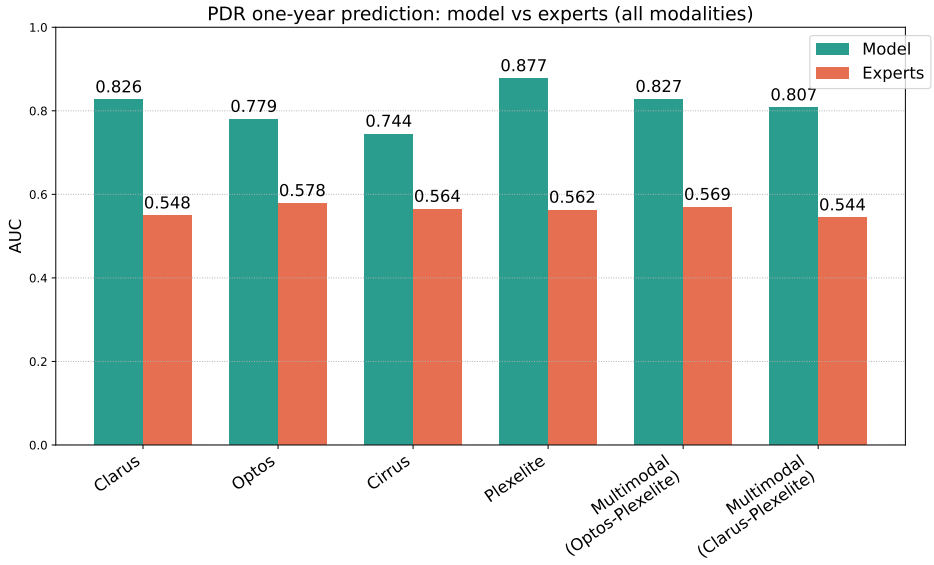


(d) PLEX Elite9000.

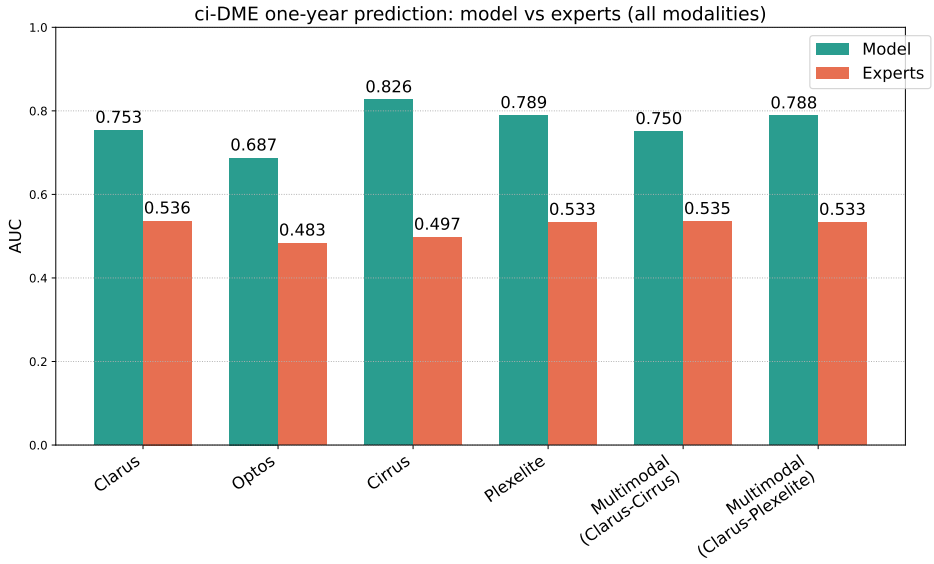
Fig. 6: Complete-case comparison with baseline risk indicators and risk stratification for one-year ci-DME prediction. Analyses are shown for **a**, CLARUS500; **b**, OPTOS; **c**, CIRRUS5000; and **d**, PLEX Elite9000. For each imaging device, the left panel compares the AUC point estimate of the image-based AI prediction with those of HbA1c, baseline diabetic macular edema status and baseline diabetic retinopathy severity. The right panel shows the observed incidence of center-involved diabetic macular edema across quartiles of the AI-derived risk score. Analyses were restricted to patient-visits with complete information for all evaluated predictors. These secondary estimates were obtained in smaller complete-case subsets and are therefore not directly comparable with the primary AUC estimates calculated in the full eligible cohort. AUC, area under the receiver operating characteristic curve; ci-DME, center-involved diabetic macular edema.

These results indicate that, within the evaluated test set, image-based AI estimates ranked patient-visits according to future complication risk more effectively than unassisted expert assessments. The comparison should nevertheless be interpreted as exploratory because the reported values represent point estimates and do not establish statistical superiority or clinical benefit.

The findings therefore support a complementary rather than substitutive role for AI. Standardized image-derived risk estimates could provide additional quantitative information to retinal specialists, while final management decisions would continue to incorporate the broader clinical context and physician judgment.



(a) Incident PDR.



(b) Incident ci-DME.

Fig. 7: Comparison of AI model and clinical expert performance for one-year prediction of vision-threatening complications. Area under the receiver operating characteristic curve for device-specific and late-fusion AI models and unassisted retinal specialist risk estimates for **a**, incident proliferative diabetic retinopathy (PDR), and **b**, incident center-involved diabetic macular edema (ci-DME), at the subsequent annual visit. Model and expert estimates were evaluated on the same held-out test cases. Values represent AUC point estimates.

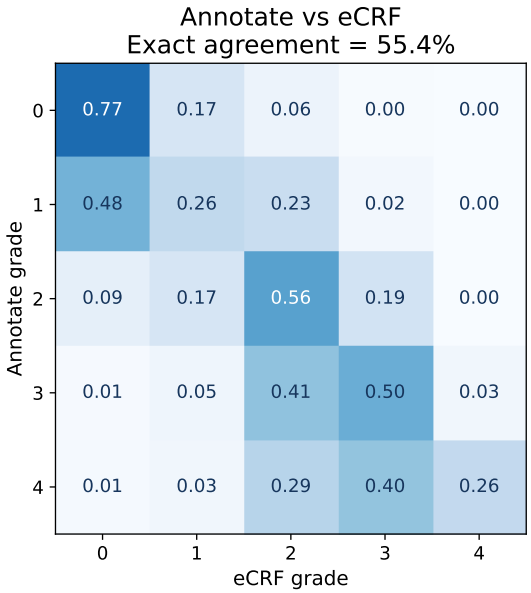


Fig. 8: Variability between local and centralized diabetic retinopathy grading. Row-normalized confusion matrix comparing DR severity grades assigned during centralized annotation with grades recorded by local investigators in the electronic case report form (eCRF). Rows represent centralized annotation grades and columns represent local eCRF grades. Values indicate the proportion of observations within each centralized annotation category. Exact agreement between the two annotation sources was 55.4%, with most disagreements occurring between neighboring severity categories.

2.9 Impact of annotation variability

We examined the agreement between DR severity grades recorded by local investigators in the electronic case report form (eCRF) and those assigned during centralized annotation (Fig. 8).

Exact agreement between the two annotation sources was 55.4%, indicating that local and centralized grades differed in nearly half of the evaluated observations. When the ordinal distance between severity categories was taken into account, the quadratic weighted kappa was 0.658.

Most disagreements occurred between adjacent or clinically related severity grades. In particular, observations assigned mild DR locally were frequently classified as either no DR or moderate DR during centralized annotation. Disagreements involving severe and proliferative disease were also concentrated among neighboring advanced categories rather than being uniformly distributed across the grading scale.

The observed variability is relevant to the interpretation of both cross-sectional and prognostic model performance because the reference grades used for model development and evaluation are themselves subject to clinical uncertainty. Nevertheless,

neither the exact agreement nor the quadratic weighted kappa should be interpreted as a formal upper bound on model performance.

These findings highlight the difficulty of establishing a single reference standard for DR severity in heterogeneous multicenter data and support the use of centralized annotation and ordinal performance metrics when evaluating model performance.

2.10 Robustness across patient subgroups, centers and cross-validation folds

We conducted exploratory analyses to examine whether model performance was maintained across clinically relevant patient subgroups, participating centers and cross-validation folds. These analyses were intended to characterize the internal robustness of the models within the EviRed cohort and were not designed as formal tests of subgroup equivalence, fairness or transportability to previously unseen centers.

For one-year DR severity prediction using CLARUS500 images, QWK estimates ranged from 0.802 to 0.886 across the evaluated demographic and clinical subgroups (Table 4). Performance was similar across age categories, with QWK values of 0.859, 0.842 and 0.839 among patient-visits from individuals aged 18–59, 60–74 and 75 years or older, respectively. QWK was 0.876 among female patients and 0.834 among male patients, and 0.886 and 0.832 among patients with type 1 and type 2 diabetes, respectively. Across the recorded ethnicity categories, QWK ranged from 0.802 in the African subgroup to 0.864 in the Caucasian subgroup. Confidence intervals overlapped across all evaluated subgroups and were wider in the smaller strata, particularly in the Asian subgroup, indicating greater uncertainty in these estimates. These exploratory analyses therefore do not provide evidence of systematic performance differences between subgroups.

For one-year ci-DME prediction using CIRRUS5000 images, AUC estimates were also broadly maintained across age, sex and diabetes type strata (Table 5). AUCs ranged from 0.804 to 0.890 across age groups and were similar among female and male patients, at 0.835 and 0.825, respectively. The corresponding AUCs were 0.851 for type 1 diabetes and 0.821 for type 2 diabetes.

Greater variation was observed across ethnicity categories, with AUC point estimates of 0.691, 0.778 and 0.857 in the African, Asian and Caucasian subgroups, respectively. However, confidence intervals overlapped across the evaluated subgroups and became substantially wider in strata with few incident events. In particular, the Asian subgroup included only 39 patient-visits and three ci-DME events, resulting in a 95% confidence interval of 0.415–1.000. Similarly, estimates among patients aged 75 years or older and those with type 1 diabetes were based on only five and seven events, respectively. Accordingly, apparent differences in subgroup point estimates should be interpreted cautiously and should not be considered evidence of differential model performance.

Center-level analyses showed that performance was not systematically associated with the number of eligible patient-visits available at each center (Figs. 9). For one-year DR severity prediction, the correlation between center size and QWK was weak and not statistically significant for both CLARUS500 (Pearson's $r = -0.129$, $p = 0.705$;

Table 4: **Subgroup performance for one-year prediction of diabetic retinopathy severity using CLARUS500 images.** Performance was evaluated at the patient-visit level. Values are point estimates with patient-clustered 95% confidence intervals.

Variable	Subgroup	Patient-visits	QWK	95% CI
Age	18–59 years	575	0.859	(0.822-0.887)
	60–74 years	519	0.842	(0.792-0.881)
	≥75 years	135	0.839	(0.738-0.896)
Sex	Female	508	0.876	(0.844-0.902)
	Male	721	0.834	(0.798-0.867)
Diabetes type	Type 1	318	0.886	(0.844-0.918)
	Type 2	896	0.832	(0.800-0.861)
Ethnicity	African	293	0.802	(0.739-0.854)
	Asian	65	0.840	(0.676-0.923)
	Caucasian	871	0.864	(0.835-0.887)

Spearman's $\rho = -0.200$, $p = 0.555$) and OPTOS (Pearson's $r = 0.091$, $p = 0.802$; Spearman's $\rho = 0.212$, $p = 0.556$).

Similarly, for ci-DME prediction using PLEX Elite9000, center-level AUC was not associated with the number of eligible patient-visits (Pearson's $r = -0.014$, $p = 0.972$; Spearman's $\rho = 0.017$, $p = 0.966$). These analyses provide no evidence that higher performance was restricted to centers contributing larger samples. Nevertheless, performance varied between centers, particularly for the ci-DME endpoint. This residual heterogeneity may reflect differences in event prevalence, patient characteristics, acquisition patterns, image quality or annotation practices. The center-level analyses therefore support robustness to center size but do not establish uniform performance across centers.

We further examined whether ci-DME prediction was driven primarily by the presence of non-central DME at the index visit. Using CIRRUS5000 images, the overall AUC was 0.826 across 590 patient-visits, including 52 incident ci-DME events (Table 6). Among patient-visits without any DME at the index visit, the model achieved an AUC of 0.804 based on 521 patient-visits and 32 events. Among patient-visits with non-central DME at the index visit, the AUC was 0.741 based on 69 patient-visits and 20 events. The persistence of discrimination within both baseline strata indicates that the model did not rely exclusively on baseline DME status to rank future ci-DME risk.

Finally, model performance was examined across cross-validation folds. For cross-sectional DR severity classification, mean QWK values were 0.889 ± 0.005 for CLARUS500, 0.871 ± 0.011 for OPTOS, 0.624 ± 0.018 for PLEX Elite9000 and

Table 5: Subgroup performance for one-year prediction of incident center-involved diabetic macular edema using Cirrus images. Values are exploratory AUC point estimates with patient-clusteredc95% confidence intervals.

Variable	Subgroup	Patient-visits	Events	AUC	95% CI
Age	18–59 years	279	27	0.804	(0.738-0.868)
	60–74 years	244	20	0.831	(0.716-0.928)
	≥75 years	67	5	0.890	(0.793-0.968)
Sex	Female	220	17	0.835	(0.754-0.906)
	Male	370	35	0.825	(0.753-0.888)
Diabetes type	Type 1	165	7	0.851	(0.758-0.933)
	Type 2	410	43	0.821	(0.757-0.876)
Ethnicity	African	74	13	0.691	(0.525-0.828)
	Asian	39	3	0.778	(0.415-1.000)
	Caucasian	477	36	0.857	(0.791-0.908)

Table 6: Sensitivity analysis according to diabetic macular edema status at the index visit. One-year ci-DME prediction performance using CIRRUS5000 images is reported overall and separately for patient-visits without DME and with non-central DME at the index visit.

Baseline status	Patient-visits	Events	AUC	95% CI
Overall	590	52	0.826	(0.770–0.878)
No DME	521	32	0.804	(0.729-0.870)
Non-central DME	69	20	0.741	(0.600-0.864)

0.507 ± 0.022 for CIRRUS5000. For one-year severity prediction, the corresponding values were 0.873 ± 0.013, 0.843 ± 0.017, 0.654 ± 0.025 and 0.493 ± 0.036, respectively.

For one-year ci-DME prediction, mean AUCs across folds were 0.793 ± 0.007 for CLARUS500, 0.760 ± 0.028 for OPTOS, 0.834 ± 0.015 for PLEX Elite9000 and 0.846 ± 0.013 for CIRRUS5000 (Table 7). The relatively small standard deviations of the primary metrics indicate that the observed ranking of the device-specific models was not dependent on a single cross-validation fold.

Together, these analyses indicate that the principal performance patterns were broadly reproduced across cross-validation folds and several clinically relevant patient strata. They also identify residual subgroup- and center-level variation that warrants further investigation. These findings support the internal robustness of the models

Table 7: Performance across cross-validation folds. Values are reported as mean ± standard deviation across folds. QWK, quadratic weighted kappa; AUC, area under the receiver operating characteristic curve; ci-DME, center-involved diabetic macular edema; PDR, proliferative diabetic retinopathy.

Imaging configuration	Cross-sectional QWK	One-year severity QWK	ci-DME AUC	PDR AUC
CLARUS500	0.889 ± 0.005	0.873 ± 0.013	0.793 ± 0.007	0.847 ± 0.040
OPTOS	0.871 ± 0.011	0.843 ± 0.017	0.760 ± 0.028	0.796 ± 0.050
CIRRUS5000	0.507 ± 0.022	0.493 ± 0.036	0.846 ± 0.013	0.743 ± 0.057
PLEX Elite9000	0.624 ± 0.018	0.654 ± 0.025	0.834 ± 0.015	0.775 ± 0.048

within the EviRed cohort but do not demonstrate equivalent performance across demographic groups or generalizability to clinical centers not represented during model development.

3 Discussion

In this prospective multicenter study, retinal imaging acquired during a single routine-care visit enabled the one-year prediction of DR severity and incident vision-threatening complications. The imaging modality providing the highest performance depended on the clinical endpoint: UWF-CFP was most informative for contemporaneous and future DR severity, whereas OCT-A yielded the highest discrimination for ci-DME and PDR. Image-derived estimates also separated patient-visits into groups with markedly different observed ci-DME incidence and achieved higher AUC point estimates than unassisted expert risk assessments. Together, these findings establish the feasibility of complementing static DR grading with quantitative short-term risk stratification.

A central finding was the limited reduction in UWF-CFP performance between contemporaneous severity classification and prediction at the subsequent annual visit. This indicates that a substantial component of short-term prognostic information is already present in retinal images acquired at the index visit. The models therefore extend the information available from conventional grading by estimating a future disease state rather than describing only the current retinal phenotype. These results do not imply that the models reconstruct individual disease trajectories or identify causal mechanisms of progression. Instead, they demonstrate that a single imaging assessment can provide prognostic information over a clinically actionable one-year horizon.

Previous studies have established that retinal imaging contains prognostic information for future diabetic retinopathy progression. Deep learning models based on color fundus photography have predicted subsequent DR worsening over fixed or variable prediction horizons, ranging from controlled clinical-trial populations to large real-world screening cohorts [9–11, 17, 18]. Previous OCT-A studies have independently shown that retinal microvascular features are associated with subsequent DR

progression and development of diabetic macular edema [19–21]. More recently, multimodal imaging and physiological biomarkers have also been evaluated for short-term prediction of vision-threatening complications [22].

The present study complements this literature by evaluating multiple UWF-CFP and OCT/OCT-A device ecosystems within the same prospective multicenter cohort and by separately predicting incident ci-DME and PDR under heterogeneous routine-care acquisition patterns. Rather than assuming that multimodal fusion uniformly improves discrimination, our results further show that the most informative imaging modality depends on the predicted clinical endpoint.

The outcome-dependent performance of the imaging modalities is clinically and methodologically relevant. UWF-CFP directly depicts the retinal lesions used to define conventional DR severity categories, which may explain its stronger performance for both current and future severity assessment. By contrast, OCT-A captures retinal structure, edema and vascular alterations that may be more directly associated with the development of ci-DME and proliferative disease. The results therefore do not support the existence of a universally optimal retinal imaging modality. Rather, they suggest that imaging should be selected and interpreted according to the complication or disease dimension being predicted.

Late fusion did not systematically outperform the strongest device-specific model. Its principal contribution was therefore operational rather than a uniform increase in discrimination. When several modalities were available, their predictions could be combined within a common framework; when only one modality had been acquired, a device-specific prediction could still be generated without requiring complete multimodal acquisition or image imputation. This flexibility is particularly relevant in routine care, where UWF-CFP and OCT-A were unavailable for 23.2% and 32.5% of observations, respectively. The value of multimodality should therefore be considered not only in terms of gains over the best unimodal model, but also in terms of coverage across heterogeneous and incomplete acquisition patterns.

The secondary complete-case analyses provided additional context for ci-DME prediction. Among patient-visits with complete information for the AI-derived estimate, HbA1c, baseline DME status and baseline DR severity, image-based predictions achieved higher AUC point estimates than each individual baseline indicator across all four imaging devices. Observed ci-DME incidence also increased progressively across quartiles of the AI-derived risk score, with the highest-risk quartile showing substantially greater event rates than the lowest-risk quartile. This gradient provides a clinically interpretable representation of risk separation beyond the global AUC.

These complete-case findings should nevertheless be interpreted cautiously. The analyses involved smaller and potentially selected subsets of the primary cohort, and their AUC estimates are therefore not directly comparable with those from the main eligible populations. Moreover, risk quartiles demonstrate stratification rather than calibration: they show that the models ranked patient-visits according to observed risk, but not that the predicted values represented accurate absolute probabilities. Comparisons with individual clinical indicators also do not establish incremental value beyond a combined multivariable clinical model. Demonstrating such value will require direct comparison of prespecified clinical and clinical-plus-imaging models on the same observations.

The comparison with retinal specialists further supports a potential complementary role for AI-derived risk assessment. Across the evaluated imaging configurations, the models achieved higher AUC point estimates than unassisted expert estimates for both ci-DME and PDR. This suggests that standardized image-derived scores may capture prognostic information that is difficult to estimate consistently through unaided clinical assessment. However, this exploratory comparison does not establish statistical superiority, improved clinical decisions or better patient outcomes. The intended role of the models is therefore to complement, rather than replace, specialist judgment by providing a reproducible quantitative estimate alongside the broader clinical context.

The robustness analyses yielded a more nuanced picture than a single overall performance estimate. The principal performance patterns were broadly maintained across cross-validation folds and several age, sex and diabetes-type strata. Center-level performance was not significantly associated with the number of eligible patient-visits contributed by each center, suggesting that the results were not restricted to the largest sites. Nevertheless, residual variation remained across centers and recorded ethnicity groups, particularly for OCT-A and ci-DME prediction. Several subgroup estimates were also based on few events. These analyses support internal robustness within the EviRed cohort, but they do not demonstrate subgroup equivalence, fairness or transportability to clinical centers that were not represented during model development and evaluation.

Variability in clinical annotation provides an important context for interpreting the severity results. Exact agreement between local and centralized grading was 55.4%, while quadratic weighted kappa was 0.658, with most disagreements occurring between adjacent severity categories. This pattern illustrates the uncertainty inherent in assigning discrete grades to a continuous and heterogeneous disease process. Similar inter-observer variability in DR severity grading has been reported previously, including for wide-field fundus photography [23]. Annotation agreement should not be interpreted as a formal upper bound on model performance, particularly because the human agreement analysis and model evaluation used different grading schemes. It nevertheless reinforces the importance of centralized annotation, ordinal evaluation metrics and sensitivity analyses in prognostic studies based on real-world clinical labels.

These findings have potential implications for risk-adapted DR management. The identification of patient-visits at increased short-term risk of ci-DME or PDR could support closer surveillance, additional imaging, or earlier specialist review. Conversely, patients with consistently low estimated risk might ultimately be considered for less intensive follow-up. Such a risk-adapted strategy would extend current follow-up approaches, which remain primarily guided by the presence and severity of diabetic retinopathy [24]. Such applications remain hypothetical. Discrimination alone is insufficient to determine whether model-guided care would improve visual outcomes, safety or resource allocation. Before clinical use, the models require external validation, assessment of calibration, prespecified operating thresholds, decision-curve analysis [25], and prospective evaluation of their effect on clinical decisions and patient outcomes.

This study has several strengths. It was based on a large prospective cohort recruited across multiple French centers and included heterogeneous UWF-CFP and OCT-A devices used in routine practice. The analyses addressed both ordinal disease severity and clinically actionable incident complications. Prediction of ci-DME and PDR was restricted to patient-visits without the corresponding complication at the index visit, thereby focusing on new events. Development and test partitions were separated at the patient level, and predictions were aggregated at the patient-visit level to reflect the bilateral nature of diabetic retinal disease. The principal estimates were also accompanied by confidence intervals and complemented by subgroup, center-level and sensitivity analyses.

Several limitations should be acknowledged. First, despite its multicenter design, the evaluation remained internal because participating centers could contribute patients to both the development and test partitions. Validation in geographically, ethnically and operationally independent cohorts is therefore required. Second, the number of incident PDR events was low, resulting in wide confidence intervals and limiting comparisons between imaging configurations. Third, missing imaging modalities were common and may not have been missing at random. Although the framework remained operational with partial imaging availability, acquisition patterns may have introduced selection bias.

Fourth, severity and outcome labels were based on clinical annotations that exhibited substantial variability. Fifth, the complete-case analyses involved smaller selected subsets and should be regarded as sensitivity analyses. Sixth, clinical variables were not used as model inputs. They were retained to contextualize the image-derived estimates in the clinician-facing report, but their potential incremental predictive contribution remains to be evaluated. Seventh, the expert comparison was exploratory and did not assess decisions made with and without AI assistance. Finally, the current framework predicts the outcome at the subsequent annual visit from a single index visit. It does not explicitly model sequences of repeated visits, treatment exposure, variable follow-up intervals or time-varying clinical covariates.

Future work should prioritize external and prospective validation, with particular attention to calibration across devices, centers and patient subgroups, together with transparent reporting according to current prediction-model guidance [26]. The incremental value of combining retinal imaging with systemic and ophthalmic clinical variables should be evaluated through prespecified multivariable comparisons. Clinically meaningful thresholds should be selected using development data and assessed in independent cohorts using sensitivity, specificity, predictive values and decision-curve analysis. Models that incorporate repeated visits, treatment history and variable prediction horizons may provide a more complete representation of disease evolution, but their added value should be demonstrated against strong single-visit baselines.

In conclusion, retinal images acquired during routine care contained information associated with DR severity and with the development of vision-threatening complications at the subsequent annual visit. The modality providing the highest performance depended on the predicted outcome, while late fusion primarily extended prediction across heterogeneous and incomplete imaging configurations. Image-derived estimates also separated patient-visits into clinically distinct ci-DME risk groups and achieved

higher discrimination than unassisted expert risk assessments. These findings support the potential of image-based AI to complement static DR grading with individualized short-term risk stratification. Prospective external evaluation is now required to determine whether these predictions can safely and effectively support risk-adapted diabetic retinopathy care.

4 Methods

4.1 Study design and population

This study was conducted within the EviRed project, a national multicenter prospective cohort designed to investigate the progression of DR under real-world clinical conditions [27].

Patients with diabetes were recruited from ophthalmology and diabetology departments across France and followed longitudinally as part of routine clinical care. Recruitment began in December 2020 and follow-up extended over multiple years depending on the time of inclusion.

The study included adult patients with type 1 or type 2 diabetes, as well as other forms of diabetes. Key exclusion criteria included absence of meaningful risk of progression, ungradable imaging in both eyes, and advanced bilateral disease precluding prediction of disease evolution.

Data were collected at baseline and during follow-up visits according to the study protocol. Patients were followed with annual visits and additional visits when clinically indicated.

All data were centralized on a common platform and independently reviewed in a dedicated reading center. Annotations were performed by trained graders with adjudication procedures to ensure consistency.

The study protocol was approved by the appropriate ethics committee, and all participants provided written informed consent.

4.2 Ethics approval

All procedures were conducted in accordance with the Declaration of Helsinki. The study protocol was approved by the French South-West and Overseas Ethics Committee 4 on 28 August 2020 (CPP2020-07-060b/2020-A01725-34/20.06.16.41433). All participants provided written informed consent prior to inclusion in the study. The study was registered at ClinicalTrials.gov (identifier: NCT04624737).

4.3 Data collection and modalities

At each visit, multimodal data were collected as part of routine clinical care, including retinal imaging and systemic clinical variables.

Retinal imaging was performed using multiple devices across participating centers, reflecting real-world clinical practice. UWF-CFP was acquired using CLARUS500 (Carl Zeiss Meditec Inc. , Dublin CA) and OPTOS systems (OPTOS plc), while OCT-A was obtained using PLEX Elite9000 9000 and CIRRUS5000 HD-OCT (Carl Zeiss Meditec Inc. , Dublin CA).

Although additional devices were used in some centers, they were not included in the present analysis due to insufficient sample size or technical limitations preventing standardized data extraction. The selected devices correspond to those with sufficient data availability.

No constraints were imposed on acquisition protocols beyond standard clinical practice, with the exception of OCT-A imaging, for which the field of view was partially standardized. Macula-centered 6×6 mm scans were required when available, corresponding to a common acquisition setting across devices such as CIRRUS5000 and PLEX systems. Larger fields of view (e.g., up to 15×15 mm for PLEX Elite9000 systems) were included when available. This approach preserved real-world heterogeneity while ensuring a minimum level of comparability across OCT-A data.

Systemic clinical variables covered a broad spectrum of patient-level information, including demographic data, diabetes characteristics, systemic comorbidities, prior complications, treatments, and detailed clinical measurements, as well as ophthalmological history and examination findings. These variables capture multiple dimensions of disease burden, ranging from metabolic control and cardiovascular status to ocular structure and function. Patient-reported outcomes were additionally collected through standardized health questionnaires. The full list of tabular features is detailed elsewhere [27].

Data were collected longitudinally, with repeated measurements over time for each patient. The number and timing of visits varied across patients, resulting in irregular follow-up intervals and incomplete modality availability.

Prior to analysis, imaging and clinical data were harmonized and linked at the eye and visit levels, enabling multimodal and longitudinal modeling.

Missing data were not artificially imputed at the data collection stage and were handled within the modeling framework.

4.4 Outcome definition

The primary outcomes of this study were the occurrence of vision-threatening DR complications within a one-year horizon, defined as PDR and ci-DME.

PDR was defined based on centralized grading of retinal imaging and corresponding clinical annotations, according to the ICDR classification.

ci-DME was assessed on structural OCT B-scans derived from OCT-A acquisitions, in accordance with standard clinical definitions.

For both outcomes, labels were derived from centralized reading and/or validated clinical annotations. When discrepancies occurred, adjudication procedures were applied to determine the final label.

Prediction tasks were defined at the visit level. For each visit at time t, the outcome was defined as the occurrence of PDR or ci-DME within the subsequent 12 months.

To ensure clinical relevance and avoid label leakage, only patients at risk at baseline were included in each task. Specifically, PDR prediction was restricted to eyes without PDR at time t, and ci-DME prediction to eyes without ci-DME at time t.

In cases where multiple visits were available within the prediction horizon, the earliest occurrence of the outcome was retained. Visits without sufficient follow-up to determine outcome status at $t + 1$ were excluded from analysis.

Two other tasks were evaluated. First, cross-sectional DR severity classification used retinal images acquired at visit t to classify the operational EviRed severity grade recorded at the same visit.

Second, one-year DR severity prediction used retinal images acquired at visit t to predict the reference severity grade at the subsequent eligible annual visit $t + 1$.

4.5 Preprocessing

UWF-CFP images underwent device-specific square center cropping followed by a common resizing and normalization procedure. CLARUS500 images were cropped to the largest centered square permitted by the original image dimensions, whereas OPTOS images were cropped to a centered 2600×2600-pixel region. Images from both devices were subsequently resized to 1024×1024 pixels, converted to tensors and normalized using the ImageNet channel means and standard deviations. The same preprocessing procedure was used across all prediction tasks for a given device.

For OCT-A, the volumetric flow acquisition and the associated retinal layer segmentations were used to generate three complementary two-dimensional representations for each acquisition. First, a superficial vascular plexus en-face image was obtained by applying maximum-intensity projection to the flow signal between the internal limiting membrane (ILM) and the inner plexiform layer (IPL). Second, a full-retina flow projection was generated between the ILM and a boundary located $70\,\mu$m anterior to the fitted retinal pigment epithelium boundary. Third, the corresponding scanning laser ophthalmoscopy localizer image was retained to provide complementary structural context.

This representation was chosen to balance computational efficiency and information content, while remaining compatible with standard deep neural network architectures. Alternative representations based on volumetric inputs (e.g., two-channel structure and flow volumes) were also explored, but showed lower predictive performance in preliminary experiments and were less suitable for clinical deployment due to increased computational complexity and data handling constraints.

4.6 Model architecture

Separate image-based models were developed for each imaging device and prediction task. The final architecture was selected within the development partition and could therefore differ between devices and outcomes. Clinical variables were not used as model inputs.

For UWF-CFP, each image was processed using a convolutional neural network implemented with the timm library. For contemporaneous DR severity classification, a ResNeXt-50 32×4d backbone was used for CLARUS500 and a ResNet-50 backbone for OPTOS. For one-year DR severity prediction, the selected backbones were ResNet-50 for CLARUS500 and ResNeXt-50 32×4d for OPTOS. For one-year ci-DME and PDR prediction, ResNeXt-50 32×4d was used for both CLARUS500 and OPTOS. All UWF-CFP models received three-channel images with an input resolution of 1024×1024 pixels.

For OCT-A, each acquisition was represented by three complementary two-dimensional images: a full-retina flow projection, a superficial vascular plexus projection and the corresponding localizer image. The three representations were processed by parallel feature extractors with the same backbone architecture but independently learned parameters.

For contemporaneous and one-year DR severity prediction, SE-ResNet-50 backbones were used for both CIRRUS5000 and PLEX Elite9000. The resulting feature vectors from the three OCT-A representations were converted into three tokens and concatenated with a learnable classification token. Positional embeddings were added, and the tokens were fused using a Transformer encoder. The final prediction was produced from the fused classification token.

For one-year ci-DME prediction, each OCT-A representation was processed using a ViT-Base/16 backbone. Features from the three representations were subsequently fused using the same token-based Transformer architecture. CIRRUS5000 models used an input resolution of 245×245 pixels, whereas PLEX Elite9000 models used an input resolution of 500×500 pixels.

All backbones were initialized from pretrained weights during model development and fine-tuned for the corresponding EviRed prediction task. The ordinal severity models produced six output logits corresponding to the operational EviRed grading categories. The binary ci-DME and PDR models produced two output logits corresponding to the absence or occurrence of the respective incident complication.

4.7 Training and evaluation protocol

Participants were divided at the patient level into a development partition comprising 65% of the cohort and a held-out test partition comprising the remaining 35%. All images, eyes, visits and visit-to-visit transitions from a given participant were assigned to the same partition, thereby preventing patient-level information leakage.

Within the development partition, five-fold cross-validation was performed at the patient level. For each fold, the model was trained on four folds and evaluated on the remaining fold. This procedure yielded five independently trained models for each imaging device and prediction task.

Model architecture, preprocessing choices and training hyperparameters were selected using performance within the development partition. For the ordinal DR severity tasks, model selection was based primarily on validation quadratic weighted kappa. For the binary ci-DME and PDR prediction tasks, selection was based primarily on validation AUC. The held-out test partition was not used for architecture selection, hyperparameter optimization or fusion development.

At inference, the predictions generated by the five fold-specific models were averaged to obtain the final device-specific prediction. When late fusion was evaluated, it was applied to the resulting device-level predictions rather than to the raw images or intermediate feature representations.

Final performance was assessed on the held-out patient test partition. No participant contributed data to both the development and test partitions. Because

participating centers could contribute patients to both partitions, this evaluation represents a multicenter internal held-out evaluation rather than external validation in previously unseen centers.

Preprocessing and training procedures were prespecified for each imaging device and applied consistently across the corresponding prediction tasks. Model comparisons were conducted under the same patient-level data partitioning and evaluation framework.

4.8 Statistical analysis

All performance analyses were conducted at the patient-visit level. Because individual participants could contribute multiple eligible visits, statistical uncertainty was estimated while accounting for within-patient dependence.

For the ordinal DR severity tasks, the primary performance metric was quadratic weighted kappa (QWK), which accounts for the ordered nature of the severity categories and assigns greater weight to larger classification errors. Exact classification accuracy was reported as a complementary metric.

For the binary prediction of incident ci-DME and PDR, discrimination was primarily assessed using the area under the receiver operating characteristic curve (AUC). Receiver operating characteristic curves were generated from the continuous patient-visit-level risk estimates. Additional binary classification metrics were reported where applicable.

During model development, performance across the five patient-level cross-validation folds was summarized as the mean ± standard deviation. These values describe variability across development folds and were reported separately from the final performance estimates obtained on the held-out test partition.

On the held-out test partition, QWK, accuracy and AUC estimates were accompanied by 95% confidence intervals derived using a non-parametric patient-clustered bootstrap with 2000 resamples. Patients were sampled with replacement, and all eligible visits belonging to each selected patient were retained within the corresponding bootstrap sample. Confidence intervals were calculated using the percentile method. Bootstrap samples in which a binary outcome contained only one class were excluded from the corresponding AUC calculation.

Comparisons between imaging devices and fusion configurations were performed within their respective eligible test populations. Because device availability differed across patient-visits, performance estimates obtained from different device-specific cohorts were not interpreted as formal paired comparisons unless explicitly stated.

For the complete-case comparison of ci-DME predictors, AUCs for the AI-derived risk estimate, HbA1c, baseline DME status and baseline DR severity were calculated within the same device-specific complete-case subset. Observed ci-DME incidence was additionally summarized across quartiles of the AI-derived risk score. These analyses were considered secondary and exploratory.

Subgroup and center-level analyses were also exploratory. Subgroup performance was reported together with the number of patient-visits and, for binary outcomes, the number of incident events. Center-level associations between performance and sample size were assessed using Pearson and Spearman correlation coefficients. No

subgroup analysis was interpreted as a formal test of equivalence, fairness or external generalizability.

All statistical analyses were performed using Python version **3.12.7** and scikit-learn version **1.5.1**, together with NumPy version **1.26.4**, SciPy version **1.13.1** and pandas version **2.2.2**.

4.9 Robustness and sensitivity analyses

A series of exploratory analyses was conducted to characterize model performance across patient subgroups, participating centers, cross-validation folds and alternative analytical subsets.

For one-year DR severity prediction using CLARUS500 images, performance was evaluated across age, sex, diabetes type and recorded ethnicity categories. For one-year ci-DME prediction using CIRRUS5000 images, discrimination was assessed across the same patient characteristics. The number of eligible patient-visits was reported for each subgroup, together with the number of incident events for the binary ci-DME outcome. These analyses were descriptive and were not designed as formal tests of subgroup equivalence or algorithmic fairness.

Center-level performance was assessed separately for each imaging device. Center-specific QWK values for the ordinal severity task and AUC values for the binary complication task were plotted against the number of eligible patient-visits contributed by each center. Pearson and Spearman correlation coefficients were calculated to examine whether model performance was associated with center sample size. Because participating centers could contribute patients to both the development and held-out test partitions, these analyses represent multicenter internal evaluation rather than validation in previously unseen centers.

Performance variability during model development was summarized across the five predefined patient-level cross-validation folds using the mean and standard deviation of the primary metric. These analyses assessed stability across the development folds but did not constitute repeated training across independently generated data partitions or multiple random initializations.

The contribution of multimodal prediction was examined by comparing late-fusion configurations with the corresponding device-specific models. These comparisons were interpreted descriptively because imaging availability differed across patient-visits and the eligible populations were not necessarily identical across configurations.

To prevent patient-level and temporal information leakage, all images, eyes, visits and visit-to-visit transitions from the same participant were assigned to a single data partition. For the prognostic tasks, models used only images available at the index visit t, whereas the reference outcome was defined from the subsequent eligible visit.

Additional sensitivity analyses were conducted for ci-DME prediction. Model performance was evaluated separately among patient-visits without DME and with non-central DME at the index visit. A secondary complete-case analysis was also performed among patient-visits with complete information for the AI-derived risk estimate, HbA1c, baseline DME status and baseline DR severity. These complete-case analyses were interpreted separately from the primary evaluation because the additional restriction reduced the sample size and may have introduced selection bias.

Annotation variability was assessed by comparing local DR severity grades recorded in the electronic case report form with centralized reading-center grades. Exact agreement and quadratic weighted kappa were calculated among observations with both annotations available. This analysis was used to characterize variability in clinical grading and was not interpreted as a formal upper bound on model performance.

Collectively, these analyses were intended to examine the internal robustness of the predictive framework under the demographic, clinical and acquisition heterogeneity represented in the EviRed cohort. They do not establish external generalizability, subgroup equivalence or absence of center-specific effects.

Data Availability. The data used in this study are part of the EviRed cohort. Due to regulatory and contractual constraints, the dataset is not immediately publicly available. The data will be made available after an embargo period, in accordance with institutional and ethical approvals. Access to the data may be granted upon reasonable request to the corresponding author, subject to approval by the relevant authorities and data protection regulations.

Code Availability. The code developed for this study is subject to an exclusive licensing agreement with industrial partners and is therefore not publicly available.

Details on the implementation, including model architecture, training procedures, and software dependencies, are provided in the Methods section to ensure reproducibility. The models were implemented using PyTorch Lightning and the timm library for deep learning, together with standard scientific Python libraries.

Additional information may be made available from the corresponding author upon reasonable request, subject to applicable intellectual property and contractual restrictions.

Acknowledgements. This work was funded in part by the French National Research Agency under the "Investissement d'Avenir" program (ANR-18-RHUS-0008 - EviRed project).

Author contributions. A.C., G.Q., P.M., E.V., and A.L.G conceived and designed the study. M.E.H.D., G.Q , P.Z., and A.R. developed the models and conducted the experiments. G.Q., M.E.H.D., and P.Z. analyzed the results and prepared the figures. M.T. and M.L. contributed to methodological development. A.C., G.Q and P.M interpreted the results. A.C., P.M., H.L.B., and the EviRed Investigators contributed to patient recruitment, clinical data acquisition and clinical interpretation. E.V. contributed to the study design and statistical methodology. A.Z., C.L., M.G., M.B., P.D., A.D., F.P., L.B., A.L.G., S.G., A.S. and D.C. contributed to imaging resources, technical infrastructure and data management. C.L. and A.Z managed the project. G.Q., M.E.H.D., and P.Z and drafted the manuscript. All authors reviewed the manuscript.

Competing Interests. A.C. serves as a consultant to ZEISS. G.Q. serves as a consultant to Evolucare Technologies and ADCIS. P.Z. has been employed by Evolucare Technologies. P.D., M.B., A.D. and F.P. are employees of ADCIS. L.B. and A.L.G. are employees of Evolucare Technologies. S.G. and A.S. are employees of Carl Zeiss

Meditec, Inc. D.C. is an employee of Carl Zeiss Meditec AG. The remaining authors declare no competing interests.

References

[1] Yau, J. W. Y. *et al.* Global prevalence and major risk factors of diabetic retinopathy. *Diabetes Care* **35**, 556–564 (2012).

[2] Curran, K. *et al.* Global estimates on the number of people blind or visually impaired by diabetic retinopathy: A meta-analysis from 2000 to 2020. *Eye* **38**, 2047–2057 (2024).

[3] Cheung, N., Mitchell, P. & Wong, T. Y. Diabetic retinopathy. *Lancet* **376**, 124–136 (2010).

[4] Wilkinson, C. P. *et al.* Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmol* **110**, 1677–1682 (2003).

[5] American Diabetes Association Professional Practice Committee for Diabetes. 12. retinopathy, neuropathy, and foot care: Standards of care in diabetes—2026. *Diabetes Care* **49**, S261–S276 (2026).

[6] Abràmoff, M. D., Garvin, M. K. & Sonka, M. Retinal imaging and image analysis. *IEEE Rev Biomed Eng* **3**, 169–208 (2010).

[7] Elsharkawy, M. *et al.* The role of different retinal imaging modalities in predicting progression of diabetic retinopathy: A survey. *Sensors* **22**, 3490 (2022).

[8] Scheideman, A. F. *et al.* Machine learning to diagnose complications of diabetes. *J Diabetes Sci Technol* **19**, 1650–1670 (2025).

[9] Nderitu, P. *et al.* Predicting 1, 2 and 3 year emergent referable diabetic retinopathy and maculopathy using deep learning. *Commun Med* **4**, 167 (2024).

[10] Silva, P. S. *et al.* Automated machine learning for predicting diabetic retinopathy progression from ultra-widefield retinal images. *JAMA Ophthalmol* **142**, 171–177 (2024).

[11] Dai, L. *et al.* A deep learning system for predicting time to progression of diabetic retinopathy. *Nat Med* **30**, 584–594 (2024).

[12] Usman, T. M., Saheed, Y. K., Nsang, A., Ajibesin, A. & Rakshit, S. A systematic literature review of machine learning based risk prediction models for diabetic retinopathy progression. *Artif Intell Med* **143**, 102617 (2023).

[13] Zhou, J., Hu, C., Chen, J., Liu, J. & Li, X. Artificial intelligence in screening and grading diabetic eye diseases: A systematic review from algorithms to clinic.

Diabetes Obes Metab **In press** (2026).

[14] Ali, S. G. *et al.* Revolutionizing diabetic retinopathy and macular edema management: A systematic review on the transformative potential of artificial intelligence. *Vis Comput* **41**, 10109–10134 (2025).

[15] 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).

[16] Zhang, Y. *et al.* Artificial intelligence-enabled screening for diabetic retinopathy: A real-world, multicenter and prospective study. *BMJ Open Diab Res Care* **8**, e001596 (2020).

[17] Arcadu, F. *et al.* Deep learning algorithm predicts diabetic retinopathy progression in individual patients. *npj Digital Medicine* **2**, 92 (2019).

[18] Bora, A. *et al.* Predicting the risk of developing diabetic retinopathy using deep learning. *The Lancet Digital Health* **3**, e10–e19 (2021).

[19] Sun, Z. *et al.* OCT angiography metrics predict progression of diabetic retinopathy and development of diabetic macular edema: A prospective study. *Ophthalmology* **126**, 1675–1684 (2019).

[20] Ding, X. *et al.* Expanded field OCT angiography biomarkers for predicting clinically significant outcomes in non-proliferative diabetic retinopathy. *American Journal of Ophthalmology* **270**, 216–226 (2025).

[21] Yang, Q. *et al.* Using OCT angiography to predict diabetic retinopathy progression and vision decline in a multiethnic cohort. *Ophthalmology Science* **6**, 101111 (2026).

[22] Davis, C. Q., Waheed, N. K. & Brigell, M. Predicting progression to vision-threatening complications in diabetic retinopathy. *Ophthalmology Science* **5**, 100859 (2025).

[23] Srinivasan, S. *et al.* Inter-observer agreement in grading severity of diabetic retinopathy in wide-field fundus photographs. *Eye* **37**, 1231–1235 (2023).

[24] Lim, J. I. *et al.* Diabetic retinopathy preferred practice pattern. *Ophthalmology* **132**, P75–P162 (2025).

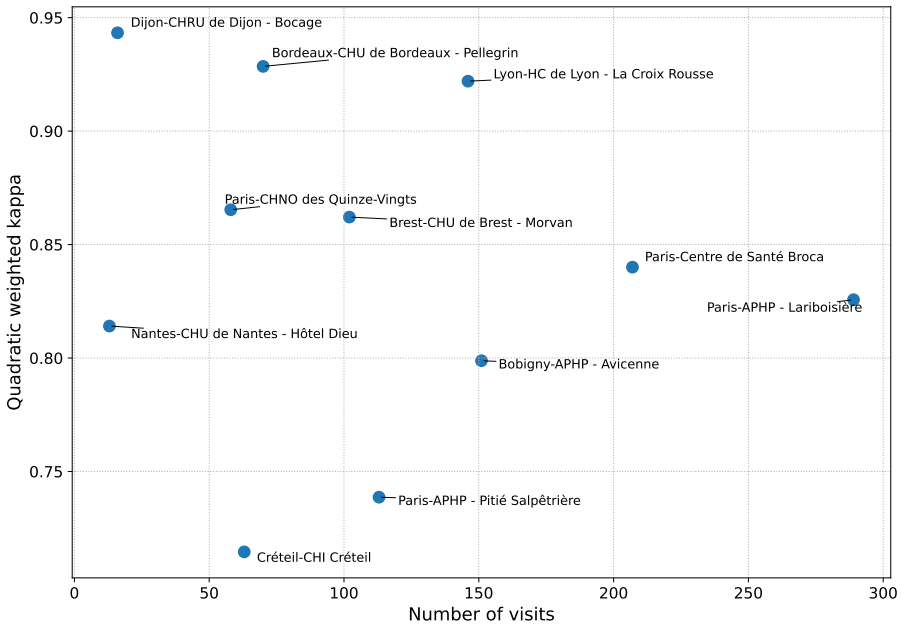
[25] Vickers, A. J. & Elkin, E. B. Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making* **26**, 565–574 (2006).

[26] 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).

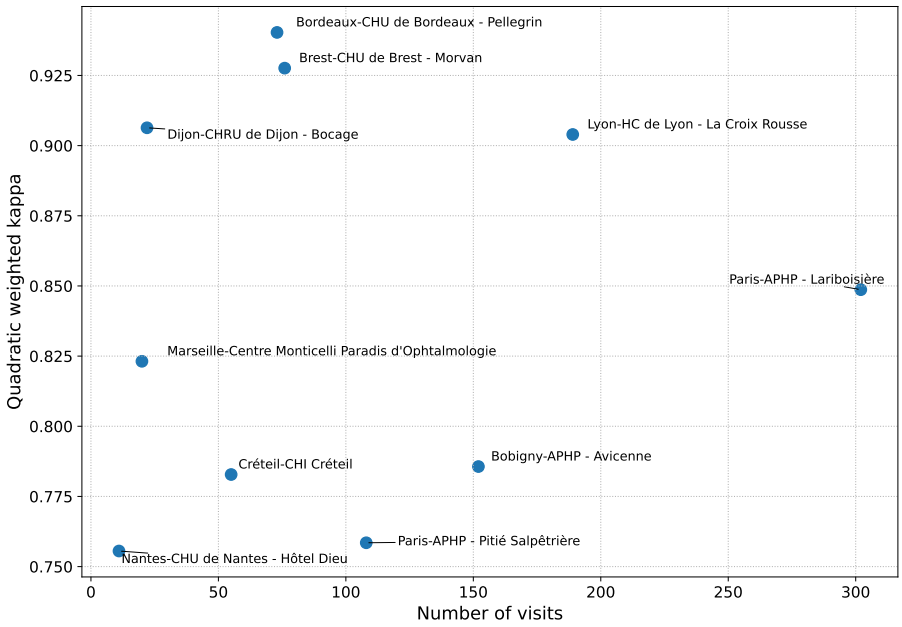
[27] Tadayoni, R. *et al.* Artificial intelligence-based prediction of diabetic retinopathy evolution (EviRed): Protocol for a prospective cohort. *BMJ Open* **14**, e084574 (2024).

EviRed investigators

Aude Couturier; Catherine Creuzot-Garcher; Jean-François Korobelnik; Laurent Kodjikian; Audrey Giocanti; Alexandra Miere; Sarah Touhami; Pascale Massin; Béatrice Cochener; Sarah Ayello-Scheer; Michel Weber; Sophie Bonnin; Pierre Gascon; Stéphanie Baillif; Jean-François Gautier; Louis Potier; Bruno Verges; Vincent Rigalleau; Philippe Moulin; Paul Valensi; Emmanuel Cosson; Christiane Ajzenberg; Agnès Hartmann; Etienne Larger; Alfred Penfornis; Véronique Kerlan; Sylvie Feldman; Camille Vatier; Samy Hadjadj; Clara Bouche; Nicolas Chevalier; Bénédicte Gaborit.



(a) One-year DR severity prediction using CLARUS500 images.



(b) One-year DR severity prediction using OPTOS images.

Fig. 9: Center-level performance for one-year prediction of diabetic retinopathy severity. Quadratic weighted kappa is plotted against the number of eligible patient-visits for each participating center for **a**, CLARUS500, and **b**, OPTOS. Each point represents one center. Pearson and Spearman correlation analyses showed no statistically significant association between center sample size and performance. These analyses were conducted within the held-out patient test set and do not constitute external validation in centers excluded from model development.