Performance of AI-based diabetic retinopathy screening is highly dependent on evaluation setting: A five-year, multi-framework study

Gwenolé Quellec

gwenole.quellec@inserm.fr

Inserm

Mathieu Lamard

University of Western Brittany

Sarah Matta

University of Western Brittany

Laurent Borderie

Evolucare Technologies

Philippe Zhang

University of Western Brittany

Paola Ortega Saborio

Evolucare Technologies

Pierre Deman

Evolucare Technologies

Alexandre Le Guilcher

Evolucare Technologies

Ali Erginay

Assistance Publique - Hôpitaux de Paris

Anna-Maria Kubin

Oulu University Hospital

Nina Hautala

Oulu University Hospital

Petri Huhtinen

Optomed

Béatrice Cochener

Centre Hospitalier Régional Universitaire de Brest

Pascale Massin

Assistance Publique - Hôpitaux de Paris

Article

Keywords: Diabetic retinopathy screening, artificial intelligence, real-world evaluation, domain shift, external validation, performance variability, medical imaging, screening programs

Posted Date: May 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9427241/v1

Additional Declarations: Competing interest reported. L.B., P.O.S., P.D. and A.L.G. are employees of Evolucare Technologies. P.H. is affiliated with Optomed and holds equity in the company. G.Q. and M.L. report that a software license related to the evaluated system has been granted to Evolucare Technologies. G.Q. has also served as a consultant for Evolucare Technologies. The remaining authors declare no competing interests.

Performance of AI-based diabetic retinopathy screening is highly dependent on evaluation setting: A five-year, multi-framework study

Gwenolé Quellec[1*], Mathieu Lamard[2,1], Sarah Matta[2,1], Laurent Borderie[3], Philippe Zhang[2,1], Paola Ortega Saborio[3], Pierre Deman[3,4], Alexandre Le Guilcher[3], Ali Erginay[5], Anna-Maria Kubin[6,7], Nina Hautala[6,7], Petri Huhtinen[8], Béatrice Cochener[9,2,1], Pascale Massin[5]

[1]LaTIM UMR 1101, Inserm, Brest, 29200, France.
[2]Université de Bretagne Occidentale, Brest, 29200, France.
[3]Evolucare Technologies, Villers-Bretonneux, 80800, France.
[4]ADCIS S.A., Saint-Contest, 14280, France.
[5]Département d'ophtalmologie, Hôpital Lariboisière, AP-HP, Paris, 75010, France.
[6]Department of Ophthalmology, Oulu University Hospital, Oulu, 90220, Finland.
[7]University of Oulu, Oulu, 90220, Finland.
[8]Optomed, Oulu, 90230, Finland.
[9]Service d'Ophtalmologie, CHU Brest, Brest, 29200, France.

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

Abstract

Despite near-perfect performance reported on benchmark datasets, the real-world behavior of artificial intelligence (AI) systems for diabetic retinopathy (DR) screening remains insufficiently characterized. Here, we report a five-year, multi-framework evaluation of a CE-marked AI system for automated detection of referable DR and diabetic macular edema from color fundus photographs. The system was initially validated on the Messidor-2 benchmark dataset and subsequently assessed across three independent evaluation settings: a large-scale masked comparative study (US Veterans Affairs), an external validation using

handheld fundus photography in Finland, and an open, large-scale comparative evaluation within the UK National Health Service.
Across these evaluation frameworks, substantial variability in observed sensitivity and specificity was found. Differences in imaging devices, population characteristics, referral definitions, and handling of ungradable images contributed to these variations. In masked settings, interpretation of comparative performance was limited by the lack of system-level attribution, while open evaluations remained sensitive to protocol design.
Taken together, these results show that performance in AI-based DR screening is not a fixed property of a system, but an emergent property of the evaluation framework. This work calls for context-aware validation strategies and more standardized evaluation protocols for clinical AI systems.

Keywords: Diabetic retinopathy screening; artificial intelligence; real-world evaluation; domain shift; external validation; performance variability; medical imaging; screening programs

1 Introduction

Artificial intelligence (AI) systems have demonstrated high diagnostic performance for diabetic retinopathy (DR) screening on benchmark datasets, with several studies reporting near-expert or even expert-level accuracy [1, 2]. These advances have supported regulatory approvals and the deployment of AI-based screening systems in clinical practice [3, 4]. However, the extent to which such performance translates across real-world settings remains insufficiently characterized. Recent meta-analyses further support acceptable performance in real-world settings, while also highlighting substantial variability across studies [5].

A growing body of evidence indicates that AI systems are sensitive to variations in data distribution, commonly referred to as domain shift [6, 7]. This includes covariate shift, corresponding to changes in the input distribution (X), as well as shifts affecting the relationship between inputs and labels (y). Under such conditions, model performance may degrade substantially when training and deployment data differ [8, 9]. In clinical settings, this issue is further compounded by differences in the definition of the target condition, often referred to as concept shift, which can be interpreted as a shift in the target variable (y), for example through variations in referral criteria or in the handling of ungradable images [10]. In retinal imaging, these sources of variability frequently co-occur, arising from differences in imaging devices (e.g., tabletop versus handheld cameras), acquisition protocols, patient populations, image quality, and clinical definitions. Together, these factors challenge the notion that performance measured on a given dataset reflects an intrinsic property of an AI system.

To address these issues, several independent evaluations of DR screening systems have been conducted in recent years, including large-scale comparative studies under both masked and open conditions [11, 12]. While these studies provide valuable insights, their interpretation remains complex due to heterogeneity in study design, evaluation protocols, and reporting practices. In particular, masked evaluations

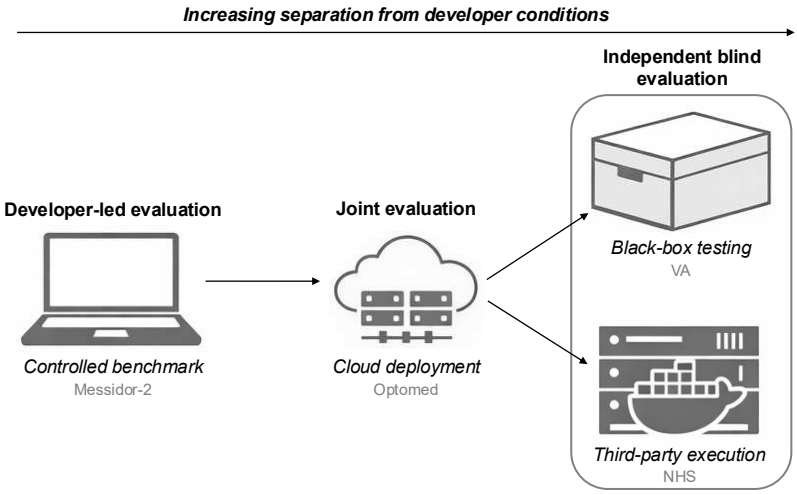

Fig. 1 Overview of evaluation workflows across settings. Schematic representation of the different evaluation environments considered in this study. From left to right, the level of separation from development conditions increases, from developer-led benchmark evaluation on a controlled dataset (Messidor-2), to collaborative evaluation on an external platform (Optomed), and finally to independent evaluation frameworks. These include masked black-box testing (PC shipped to the evaluators) with no system-level attribution (VA: US Veterans Affairs) and standardized third-party execution (Docker-based submission) under a predefined protocol (NHS: UK National Health Service). The branching structure highlights that masked and open evaluations represent distinct paradigms, reflecting different forms of independence and protocol control.

limit system-level attribution, whereas open evaluations may still depend strongly on protocol-specific choices, making cross-study comparisons difficult.

In this work, we investigate how the observed performance of an AI system for DR screening varies across distinct evaluation frameworks (see Fig. 1). We report a five-year, multi-framework evaluation of a CE-marked system across four independent settings: a controlled benchmark dataset (Messidor-2), a large-scale masked comparative study (VA: US Veterans Affairs), an external validation using handheld fundus photography in Finland [13], and an open, large-scale comparative evaluation within the UK National Health Service (NHS). Rather than treating these evaluations as independent confirmations of performance, we analyze them as complementary perspectives on how performance depends on the evaluation context.

Our results show that the same system can exhibit markedly different operating characteristics across evaluation settings. These findings suggest that performance in AI-based DR screening is not a fixed property of a system, but a function of the evaluation framework. This work highlights the need for context-aware validation strategies and more standardized evaluation protocols for clinical AI.

2 Results

Performance across four distinct evaluation frameworks is reported below and illustrated in Fig. 2, where each point corresponds to a distinct evaluation setting in the

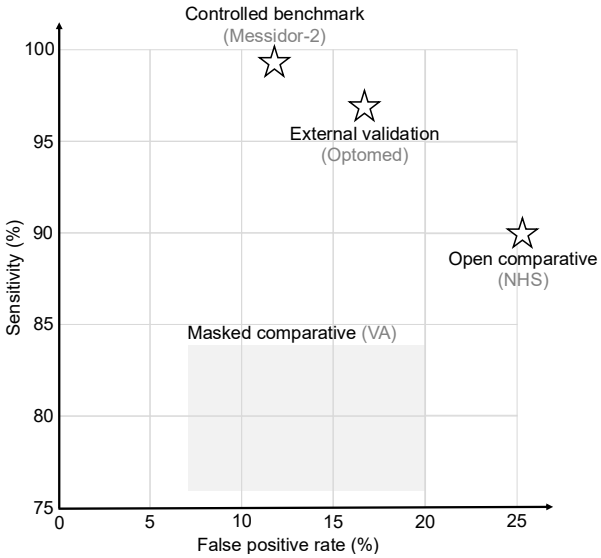

Fig. 2 Performance of the same AI system across four distinct evaluation frameworks. Each point represents the observed sensitivity and specificity for referable DR detection in a given evaluation setting: controlled benchmark (Messidor-2), external validation under acquisition shift (Optomed), and open comparative evaluation (NHS: UK National Health Service). The shaded area indicates the range of reported sensitivity and specificity across systems in the masked comparative evaluation (VA: US Veterans Affairs). Substantial variability in operating characteristics is observed across settings. This variability reflects differences in imaging devices, populations, referral definitions, and evaluation protocols, illustrating the context dependence of AI performance.

sensitivity–specificity space. The main characteristics of these evaluation settings are summarized in Table 1.

Table 1 Characteristics of the four evaluation settings.

Setting	Type	Country	Scale	Acquisition device
Messidor-2	Controlled Benchmark	France	1,748 images (874 patients)	Tabletop cameras
VA	Masked comparative	USA	311,604 images (23,724 patients)	Tabletop cameras
Optomed	External validation	Finland	1,248 images (156 patients)	Handheld camera
NHS	Open comparative	UK	∼1.2 million images (202,886 encounters)	Mixed devices

2.1 Controlled benchmark evaluation (Messidor-2)

On the Messidor-2 benchmark dataset [2, 14], the system achieved high diagnostic performance for the detection of referable DR, with an area under the receiver operating characteristic curve (AUC) of approximately 0.99 (see Table 1 of the Supplementary

Information). Two operating points were identified: a high-sensitivity setting (sensitivity = 99.0%, specificity = 88.2% — see Fig. 1 of the Supplementary Information) and a high-specificity setting (sensitivity = 95.3%, specificity = 92.1%). The high-sensitivity operating point was used for comparisons with other evaluation settings, reflecting typical screening requirements.

This evaluation was conducted under controlled conditions, with standardized image acquisition and reference annotations. As such, it provides a stable estimate of performance under benchmark settings, but does not account for variability introduced by real-world clinical environments.

2.2 Masked comparative evaluation (VA study)

In a large-scale, multi-center comparative evaluation conducted within the VA healthcare system, multiple AI-based DR screening systems were assessed under masked conditions [11]. System-level attribution of performance metrics was not disclosed. Within this framework, the evaluated system was included among the approaches selected for detailed analysis by the study investigators. However, the absence of explicit system-level results limits quantitative interpretation and prevents direct comparison with other methods.

Reported sensitivities and specificities among the systems selected for detailed analysis spanned approximately 80% to 93% and 76% to 84%, respectively. This variability further illustrates the difficulty of interpreting comparative performance under masked conditions, where differences in evaluation design and operating points cannot be fully disentangled.

This setting highlights a key limitation of masked evaluations: while they reduce bias in comparative studies, they also restrict interpretability and hinder the translation of results into actionable performance estimates.

2.3 External validation under acquisition shift (Optomed)

The system was evaluated on an independent dataset acquired at Oulu University Hospital using a handheld fundus camera under non-mydriatic conditions [13]. Compared to the benchmark setting, this evaluation involved differences in imaging device, acquisition conditions, and grading protocols.

In this context, the system achieved a sensitivity of 83.6% and a specificity of 96.8% for referable DR detection. While specificity remained high, sensitivity decreased compared to the benchmark evaluation. This performance shift is consistent with differences in image acquisition, field of view, and disease prevalence, as well as variations in grading standards.

This evaluation illustrates the impact of domain shift on observed performance, particularly when moving from controlled datasets to heterogeneous real-world acquisition conditions.

2.4 Open comparative evaluation (UK National Health Service)

The system was further evaluated in a large-scale, open comparative study conducted within the NHS, where performance metrics for all participating systems were reported transparently [12]. This evaluation framework enables direct comparison across systems under a shared protocol.

In this setting, the system achieved a sensitivity of 89.8% with a corresponding specificity of approximately 74.5%. Similar operating characteristics were observed across other evaluated systems, reflecting a common trade-off imposed by the screening protocol, which prioritizes sensitivity over specificity. This pattern suggests that the observed performance is driven not only by system characteristics, but also by the evaluation framework itself.

However, performance remained dependent on study-specific design choices, including referral definitions and handling of ungradable images. Although open evaluations improve interpretability compared to masked studies, they do not fully eliminate heterogeneity, and results remain tied to the specific evaluation protocol.

2.5 Cross-setting variability in observed performance

Across these four evaluation frameworks, the same system exhibited substantial variability in observed operating characteristics. Sensitivity and specificity varied across settings in association with differences in imaging devices, population characteristics, referral definitions, and evaluation protocols. Sensitivity ranged from approximately 80% to 99% across evaluation settings, with specificity ranging from 74.5% to 96.8%.

These evaluations should not be interpreted as independent confirmations of performance, but rather as complementary observations of how performance depends on the evaluation context. The observed variability is consistent with the combined effects of covariate shift, related to changes in image acquisition and population characteristics, and concept shift, related to differences in clinical definitions and evaluation criteria.

3 Discussion

This study provides a multi-framework evaluation of an AI system for DR screening across four independent settings spanning controlled benchmarks, masked comparative studies, external validation, and open large-scale evaluations. Rather than yielding a single estimate, these results demonstrate substantial variability in observed operating characteristics across evaluation frameworks.

A key finding is that performance should not be interpreted as an intrinsic property of the system, but as an emergent property of its interaction with the evaluation context. Differences in imaging devices, acquisition conditions, patient populations, referral definitions, and handling of ungradable images all contribute to variations in observed sensitivity and specificity, consistent with the combined effects of covariate and concept shift.

A particularly important source of concept shift arises from differences in the definition of referable DR across evaluation settings. In some screening programs, such as those used in the VA and NHS evaluations, ungradable images are considered a criterion for referral, whereas this was not the case in the benchmark dataset and in the external validation setting, where image quality was controlled or pre-filtered. More broadly, variations in referral criteria, inclusion of DME, or handling of ungradable images lead to target or label drift, effectively altering the classification task. As a result, reported performance may reflect a combination of disease detection and ancillary tasks such as image quality assessment. Differences in disease prevalence may further contribute through their impact on decision thresholds, calibration, and case mix complexity.

The comparison across evaluation settings highlights structural limitations of current evaluation practices. Benchmark datasets provide stable and reproducible estimates but fail to capture real-world variability. Masked comparative studies reduce bias but limit interpretability by preventing system-level attribution. Open comparative evaluations improve transparency, yet remain dependent on protocol-specific design choices. These frameworks therefore provide complementary but inherently partial views of system performance.

Performance variability is further influenced by the degree of independence between developers and evaluation workflows. Benchmark evaluations are typically conducted under developer control, whereas large-scale comparative studies are performed by independent third parties under predefined constraints, with external validation settings lying between these extremes. These differences reflect varying levels of control over data access, deployment conditions, and operating point selection, and should be considered when interpreting results.

Another source of variability lies in the choice of operating points. Sensitivity and specificity correspond to specific decision thresholds that may differ across settings, meaning that comparisons may reflect differences in operating regimes rather than intrinsic system performance.

More broadly, AI systems are trained on historical data but evaluated in evolving clinical environments, where imaging technologies, clinical practices, and patient populations may change over time, further contributing to discrepancies between expected and observed performance.

Taken together, these findings support a shift from viewing performance as a fixed metric to understanding it as a context-dependent property. Variability in reported performance should therefore be interpreted not as noise, but as a structural feature of AI evaluation in heterogeneous clinical settings. These findings also align with recent efforts to improve the reporting and evaluation of clinical AI systems, such as the CONSORT-AI and SPIRIT-AI guidelines [15].

These observations have implications for both development and assessment. Performance measured on a single dataset or protocol should be interpreted as a context-dependent estimate rather than an intrinsic property of a system. Evaluation strategies should explicitly account for variability across settings, and reporting should clearly specify operating points and evaluation protocols.

A related implication concerns the alignment between model design and evaluation requirements. Models are often trained to predict a single definition of referable disease, whereas evaluation criteria vary across settings. Designing models at a finer level of granularity, with outputs that can be aggregated according to different referral definitions, may improve robustness without requiring retraining.

These findings have direct clinical implications: performance should not be assumed transferable across settings, and must be established in the target population and clinical workflow prior to deployment.

This study has several limitations. The number of evaluation settings is limited, and the analysis remains descriptive. Differences in evaluation protocols could not be fully controlled, and some settings restrict access to detailed system-level results. Nevertheless, the consistency of variability across independent settings supports the robustness of the conclusions.

In conclusion, the performance of AI systems for DR screening depends critically on the evaluation framework. Recognizing and accounting for this context dependence is essential for reliable clinical deployment.

4 Methods

4.1 Study design

This study reports a retrospective, multi-framework evaluation of a CE-marked AI system for DR screening. Rather than relying on a single dataset or evaluation protocol, we analyzed the observed performance of the same system across four independent evaluation settings spanning a five-year period. These settings include a controlled benchmark dataset, a masked large-scale comparative study, an external validation under different acquisition conditions, and an open large-scale comparative evaluation. The system was not retrained or modified between evaluation settings.

4.2 AI system

The evaluated system is a CE-marked AI-based screening tool designed for automated analysis of color fundus photographs. It performs multiple tasks, including image quality assessment, eye laterality identification, detection of referable DR, detection of DME, and grading of DR severity.

The system was trained on large-scale annotated datasets derived from the OPHDIAT telemedicine screening program [16], which collects retinal images from multiple clinical centers under routine screening conditions. The training dataset included 763,848 images from 102,257 diabetic patients acquired between 2004 and 2017 under routine clinical conditions. Image grading was performed by certified ophthalmologists following standardized protocols, including quality control procedures such as double grading, and was consistent with French national guidelines for DR screening [17].

All data used for model development were collected and processed in compliance with applicable data protection regulations. The French dataset was approved by the "Commission Nationale de l'Informatique et des Libertés" (CNIL, approval #2166059), and all procedures adhered to the principles of the Declaration of Helsinki.

The system relies on ensembles of convolutional neural networks trained on these annotated datasets. Input images are preprocessed and multiple network architectures are combined to improve robustness. For each eye, predictions are computed at the image level and aggregated by selecting the image with the highest predicted severity. A detailed description of the model architectures, training procedures, and preprocessing steps is provided in the Supplementary Information.

Only the original version of the system, corresponding to the CE-marked release, was evaluated in this study. Subsequent developments were not considered, as they were not consistently evaluated across all settings.

4.3 Evaluation settings

The system was evaluated across four independent settings, each corresponding to a distinct evaluation framework differing in data characteristics, evaluation protocols, and degree of independence from the system developers.

Controlled benchmark dataset (Messidor-2). Messidor-2 is a public benchmark dataset with standardized annotations, widely used for evaluating automated DR screening systems [2, 14]. In this study, evaluation was conducted under controlled conditions by one of the system developers (G.Q.). The high-sensitivity operating point identified on this dataset was used as a reference throughout the study. *Reference standard:* International Classification of Diabetic Retinopathy [18].

Masked comparative evaluation (VA study). The VA evaluation was conducted as a large-scale, multi-center comparative study under masked conditions [11]. The software, packaged by ADCIS and Evolucare on a dedicated computer, was shipped to the University of Washington, where evaluation was performed independently. The evaluated system was included among the approaches selected for detailed analysis, but explicit system-level attribution of performance metrics was not disclosed. *Reference standard:* National Teleretinal Imaging Program Guidelines [19].

External validation under acquisition shift (Optomed). The Optomed evaluation was conducted on an independent Finnish dataset acquired using a handheld fundus camera (Optomed Aurora, Oulu, Finland) under non-mydriatic conditions [13]. Evaluation was performed by the Optomed team, with technical support from Evolucare and ADCIS. Differences in imaging device and acquisition conditions make this setting representative of external validation under acquisition shift. *Reference standard:* Finnish Current Care Guidelines [20].

Open comparative evaluation (UK National Health Service). The NHS evaluation was conducted under an open comparative framework with publicly reported performance metrics [12]. The software, packaged as a Docker container, was submitted to the evaluation team and executed under predefined conditions. While this framework enables direct comparison across systems, results remain influenced by protocol-specific choices, including referral definitions and the handling of ungradable images. *Reference standard:* UK NSC Guidance Standards for the Presence and Severity of DR [21].

These settings differ substantially in imaging devices, patient populations, referral definitions, handling of ungradable images, and evaluation workflows, providing

complementary perspectives on system performance across controlled and real-world conditions.

4.4 Outcome measures

The primary outcome measures were sensitivity and specificity for the detection of referable DR at the eye level.

Operating points were defined according to each evaluation framework. On the Messidor-2 benchmark dataset, the high-sensitivity setting was retained for cross-setting comparisons, in line with screening requirements. In other evaluation settings, operating points were either fixed by the study protocol or implicitly defined by the evaluation procedures and could not be modified.

As a result, reported sensitivity and specificity values correspond to framework-specific operating conditions and are not directly comparable across studies in terms of decision thresholds. Given the heterogeneity of evaluation frameworks, performance metrics were not pooled across studies.

4.5 Analytical approach

We adopted a descriptive and comparative approach to analyze performance variability across evaluation settings. Each evaluation was considered as a distinct observation of system behavior under specific conditions. Observed differences in sensitivity and specificity were interpreted in light of documented variations in imaging conditions, population characteristics, referral definitions, and handling of ungradable images.

Rather than assuming a fixed intrinsic performance, we analyzed how evaluation context influences observed operating characteristics. Variability across settings was interpreted as being consistent with the combined effects of covariate shift and concept shift.

Data Availability. The Messidor-2 dataset is publicly available [2, 14]. The other datasets analyzed in this study were obtained from third-party sources under specific agreements and are not publicly available. Access to these data may be subject to restrictions imposed by the corresponding data providers.

Code Availability. The source code of the AI system evaluated in this study is not publicly available due to an exclusive licensing agreement with OphtAI. The system is deployed as part of a commercial platform.

Acknowledgements. The authors gratefully acknowledge all contributors involved in the different evaluation studies, including clinical teams, technical staff, and external collaborators.

We particularly thank Aaron Lee and Jiri Fajtl for their contributions to the independent evaluation processes.

This work was funded in part by a grant by the French National Research Agency under the LabCom program (ANR-19-LCV2-0005 - ADMIRE project).

Author contributions. G.Q. and P.M. conceived and designed the study, with contributions from A.E. and B.C. G.Q. and M.L. developed the AI system.

G.Q., S.M., L.B., P.Z., P.O.S., P.D., A.L.G., A.E., A.-M.K., N.H., P.H. and P.M. contributed to the evaluation of the system across the different settings.

Clinical data collection and annotation were performed by A.E., A.-M.K., N.H., B.C., P.M., and collaborators at the participating clinical centers.

Implementation of evaluation pipelines and system deployment were carried out by L.B., P.Z., P.D., A.L.G. and P.H.

All authors contributed to data interpretation, critically revised the manuscript, and approved the final version.

Competing Interests. L.B., P.O.S., P.D. and A.L.G. are employees of Evolucare Technologies. P.H. is affiliated with Optomed and holds equity in the company.

G.Q. and M.L. report that a software license related to the evaluated system has been granted to Evolucare Technologies. G.Q. has also served as a consultant for Evolucare Technologies.

The remaining authors declare no competing interests.

References

[1] 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).

[2] Abràmoff, M. D. *et al.* Automated analysis of retinal images for detection of referable diabetic retinopathy. *JAMA Ophthalmol* **131**, 351–357 (2013).

[3] Abràmoff, 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 Digital Med* **1**, 39 (2018).

[4] Grzybowski, A. & Brona, P. Approval and Certification of Ophthalmic AI Devices in the European Union. *Ophthalmol Ther* **12**, 633–638 (2023). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10011240/.

[5] Joseph, S. *et al.* Diagnostic accuracy of artificial intelligence-based automated diabetic retinopathy screening in real-world settings: A systematic review and meta-analysis. *Am J Ophthalmol* **263**, 214–230 (2024).

[6] Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V. & Herrera, F. A unifying view on dataset shift in classification. *Pattern Recogn* **45**, 521–530 (2012).

[7] Matta, S. *et al.* A systematic review of generalization research in medical image classification. *Comput Biol Med* **183**, 109256 (2024).

[8] Koh, P. W. *et al.* WILDS: A benchmark of in-the-wild distribution shifts (2021). Proceedings of the 38th International Conference on Machine Learning (ICML).

[9] Matta, S. *et al.* Towards population-independent, multi-disease detection in fundus photographs. *Sci Rep* **13**, 11493 (2023).

[10] Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* **17**, 195 (2019).

[11] 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).

[12] Rudnicka, A. R. *et al.* Automated retinal image analysis systems to triage for grading of diabetic retinopathy: A large-scale, open-label, national screening programme in England. *Lancet Digit Health* **7** (2025).

[13] Kubin, A.-M. *et al.* Comparison of 21 artificial intelligence algorithms in automated diabetic retinopathy screening using handheld fundus camera. *Ann Med* **56**, 2352018 (2024).

[14] Decencière, E. *et al.* Feedback on a publicly distributed database: The Messidor database. *Image Anal Stereol* **33**, 231–234 (2014).

[15] Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J. & Denniston, A. K. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med* **26**, 1364–1374 (2020).

[16] Massin, P. *et al.* OPHDIAT: a telemedical network screening system for diabetic retinopathy in the Ile-de-France. *Diabetes Metab* **34**, 227–234 (2008).

[17] Massin, P. *et al.* Recommandations de l'ALFEDIAM pour le dépistage et la surveillance de la rétinopathie diabétique. *J Fr Ophtalmol* **20**, 302–310 (1997).

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

[19] Conlin, P. R., Fisch, B. M., Orcutt, J. C., Hetrick, B. J. & Darkins, A. W. Framework for a national teleretinal imaging program to screen for diabetic retinopathy in Veterans Health Administration patients. *J Rehabil Res Dev* **43**, 741–748 (2006).

[20] Summanen, P. *et al.* Update on current care guideline: Diabetic retinopathy. *Duodecim* **131**, 893–894 (2015).

[21] NHS England. Diabetic eye screening: retinal image grading criteria. https://www.gov.uk/government/publications/diabetic-eye-screening-retinal-image-grading-criteria (2025). Accessed 6 April 2026.

Supplementary Files

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

- SupplementaryMaterial.pdf