nature medicine

Article

https://doi.org/10.1038/s41591-026-04545-w

AI-based clinician decision support system for diagnosis of inherited retinal diseases: a multicenter, randomized trial

Received: 7 January 2026

Accepted: 25 June 2026

Published online: 24 July 2026

Check for updates

Huixun Jia[1,25], Bo Qian[2,25], Yanlin Qu[1,25], Jinyuan Wang[3,25], Jieqiong Chen[1,25], Tong Li[1,25], Changle Zheng[2,25], Jinu Han[4,25], Guanran Zhang[1], Yixiao Jin[3], Seungmin Lee[5], Xiang Chen[5], Haichao Chen[3], Jiayu Xu[1], Ke Xu[6], Yu Rong[7], Yi Jiang[8], Yiliu Yang[9], Wenjing Li[10], Qingge Guo[11], Ya Li[11], Yang Li[6], Jihong Wu[7], Qingjiong Zhang[8], Fang Lu[9], Wenjuan Zhuang[10], Bo Lei[11,12], Suk Ho Byeon[4], Junwon Lee[13], Adrian Smedowski[14,15,16], Marta Swierczynska[17,18], Daoqiang Zhang[2], Tyler Hyungtaek Rim[19,20], Andrzej Grzybowski[21,22], Bin Sheng[5] ✉, Tien Yin Wong[3,23,24] ✉ & Xiaodong Sun[1] ✉

The accurate and timely diagnosis of inherited retinal diseases (IRDs) represents an unmet clinical need in ophthalmology, as the current pathways rely on resource-intensive phenotyping, multidisciplinary expertise and genetic testing. Here we developed Retina4IRD, an artificial intelligence (AI)-based clinician decision support system (CDSS) that predicts 17 genotype categories from retina images. Retina4IRD uses a Vision Transformer model pretrained with RETFound. We then trained and validated Retina4IRD using multimodal data with color fundus photographs and optical coherence tomography scans from 1,843 genetically confirmed patients (3,376 eyes) across China, South Korea and Poland. The top-5 prediction accuracy was 0.904 (95% confidence interval (CI): 0.896–0.912) and 0.856 (95% CI: 0.850–0.863) for internal and external validation, respectively. We conducted a randomized controlled trial with 300 participants with suspected IRD randomized 1:1 to either Retina4IRD-assisted specialist arm or specialist-only arm. Of these, 295 participants (median age 33 years, 114 (38.6%) females) with available next-generation sequencing reports were included in the final analysis. The primary outcome was met: top-5 genetic accuracy was significantly higher in the Retina4IRD-assisted specialist arm versus the specialist-only arm (88.5% versus 67.3%, $P < 0.001$). For secondary endpoints, top-1 to top-4 accuracies all favored the Retina4IRD-assisted specialist arm, with top-1 accuracy of 37.8% versus 22.4% and top-4 accuracy of 81.8% versus 53.1%, respectively. Post hoc analyses demonstrated that, with Retina4IRD assistance, clinicians made better management decisions, and the composite downstream management score indicated significantly higher scores relative to the control group (37.7 versus 28.5, $P < 0.001$). Our study shows that Retina4IRD is a CDSS tool prior to genetic testing and aligns with clinical workflow for patients with suspected IRDs. ClinicalTrials.gov identifier: NCT06839170.

A full list of affiliations appears at the end of the paper. ✉e-mail: shengbin@sjtu.edu.cn; wongtienyin@tsinghua.edu.cn; xdsun@sjtu.edu.cn

Inherited retinal diseases (IRDs), primarily caused by monogenic mutations, are a leading cause of blindness in children and working-age individuals[1]. Although individually rare, they collectively affect more than 5 million people worldwide[2,3]. To date, more than 300 nuclear and mitochondrial genes have been associated with IRDs[4]. Precise genetic diagnosis is the critical first step for targeted treatment selection, prognostic outcomes evaluation and genetic counseling.

However, the diagnostic paradigms for IRDs face major challenges due to their heavy reliance on resource-intensive approaches, often requiring multidisciplinary teams in specialized centers and the integration of comprehensive ophthalmic evaluations, including next-generation sequencing (NGS), familial segregation analysis and, in some cases, even functional validation[3]. This complex workflow results in prolonged diagnostic timelines, high cost and limited accessibility, leaving over 40% of patients with IRDs without definitive genetic diagnoses[5,6]. These challenges are exacerbated by the global shortage of eye specialists skilled in interpreting genetic retinal disorders, particularly in resource-limited settings where misdiagnosis or diagnostic delays often exceed 5 years[7,8]. Such delays may cause potentially treatable patients to miss critical therapeutic windows for advanced treatments, particularly groundbreaking therapies such as gene therapies[9–11]. Additionally, underdiagnosis can also lead to missed opportunities for clinical trial enrollment and hinder efficient identification of target populations, potentially discouraging pharmaceutical investment in gene therapy development. Moreover, patients frequently experience changes in their diagnosis as they consult multiple clinicians over time, creating an urgent need for more accessible and cost-effective diagnostic solutions.

Although artificial intelligence (AI) has revolutionized retinal disease screening[12–15], the application of AI for IRD screening and genotype prediction remains largely underdeveloped despite preliminary investigations[16–19]. When introduced, the most advanced AI system to date, Eye2Gene, represented a transformative leap forward, with 83.9% top-5 accuracy in genetic prediction[20]. Nevertheless, at present, there are limitations that may constrain its broader clinical utility. First, Eye2Gene is reliant on imaging acquired by the Heidelberg Spectralis, a device that is currently not commonly available in primary care or nonspecialized medical institutions, thereby restricting implementation in resource-limited settings. Using more widely available fundus imaging data for prediction may represent a more promising approach. Second, the latest version of the Eye2Gene study did not include mode of inheritance and age at presentation, which could have enhanced the model's diagnostic accuracy[20], although this was included in earlier Eye2Gene work. Most crucially, there is an absence of prospective validation through randomized controlled trials (RCTs) to demonstrate its real-world applicability.

To address these gaps, we present Retina4IRD, a novel foundation-model-based AI system specifically designed for genotype prediction of IRD, which emulates the clinician reasoning process through multimodal analysis of color fundus photography (CFP), optical coherence tomography (OCT) and medical metadata. We used a Vision Transformer model pretrained with RETFound to develop Retina4IRD, which was trained and then validated on 1,843 genetically confirmed cases (3,376 eyes) from nine centers across China, South Korea and Poland. Retina4IRD generates a ranked list of candidate pathogenic genetic variants and can also generate attention heatmaps to offer decision-making visualization. To validate its clinical efficacy, we conducted a prospective multicenter RCT in 300 participants with suspected IRD, comparing the Retina4IRD-assisted specialist arm versus the specialist-only arm to evaluate the diagnostic accuracy and real-world utility of Retina4IRD as a pre-test clinician decision support system (CDSS) prior to formal genetic testing in the clinical workflow.

Results

An overview of the study design is presented in Fig. 1. This is a two-stage study. Stage 1 is model development and evaluation, involving 5,167 images and 8,901 metadata items from 1,843 genetically confirmed individuals (3,376 eyes) across nine centers in China, South Korea and Poland. Stage 2 is a randomized controlled validation trial conducted at seven centers in China to evaluate the effectiveness of Retina4IRD in real-world clinical settings. A total of 300 participants were enrolled and randomized, of whom 295 with available NGS reports were included in the final analysis of the RCT. The Retina4IRD architecture is illustrated in Fig. 2.

Table 1 summarizes the baseline characteristics across all datasets, with the training dataset comprising 751 patients (1,461 eyes) and the validation and test datasets including 238 patients (453 eyes) and 148 patients (283 eyes), respectively. External validation was performed across 3 independent cohorts from China (7 centers, 455 cases, 681 eyes), Poland (87 cases, 172 eyes) and South Korea (164 cases, 326 eyes), with detailed enrollment numbers by center provided in Supplementary Table 1. The training/validation cohorts had a median age of 39 years (range, 0–78), whereas external cohorts showed younger populations (Poland: median age 17 years; South Korea: median age 28 years). Male patients predominated overall (61.2%), with consistently higher proportions across all datasets. The median age of disease onset was 10 years, with a median disease duration of 20 years. Family history was reported in 11.5–34.4% of cases, being most prevalent in the South Korea cohort (34.4%).

Among therapeutically actionable genotypes in our China training cohort, the most frequent affected genes included *USH2A* (10.5%), *CYP4V2* (6.4%), *ABCA4* (5.1%) and *RPGR* (4.0%). In the external cohorts, the distribution of causal genes varied across countries (Table 1). In the external China cohort ($n = 455$), *USH2A* was the most frequent gene (37, 8.1%), followed by *ABCA4* (23, 5.1%), *CYP4V2* (21, 4.6%), *RPE65* (21, 4.6%) and *RPGR* (17, 3.7%). By contrast, the Poland cohort ($n = 87$) showed a different spectrum, with higher proportions of *RPGR* (11, 12.6%) and *ABCA4* (9, 10.3%) and no *CYP4V2* or *RPE65* cases observed. The South Korea cohort ($n = 164$) exhibited an intermediate pattern, with similar frequencies of *ABCA4* (8, 4.9%) and *RPGR* (8, 4.9%) and lower frequencies of *USH2A* (7, 4.3%). We also present the baseline demographic and clinical characteristics of the study participants stratified by IRD status (IRD versus non-IRD) in Supplementary Table 2. Supplementary Table 3 presents the distribution of multi-devices (CFP and OCT) across training dataset, validation dataset, internal test dataset and external test dataset.

Performance of Retina4IRD

Retina4IRD was trained using a comprehensive dataset comprising 2,243 images and 4,816 metadata item entries from 989 patients (1,914 eyes), sourced from the National Clinical Research Center of Eye Diseases. All patients underwent NGS to confirm pathogenic genes.

First, the model was evaluated on an internal test set across all 17 classes using paired CFP and OCT images combined with metadata, including negative cases, potentially curable cases that involve clinical trials or approved therapy and incurable cases without any current clinical trials or approved therapy. Performance metrics including patient counts, precision and recall for each class are shown alongside the confusion matrix, and the precision exceeded 0.8 for all potentially curable genotypes except *RPE65* and *CNGA1*; detailed results are provided in Fig. 3a and Supplementary Table 4. Second, the model was evaluated on an external test set of 455 participants (681 eyes) from seven centers in China across 17 classes using paired CFP and OCT images combined with metadata, including 42 negative cases (59 eyes), 211 potentially curable cases (312 eyes) and 202 incurable cases (310 eyes). Precision exceeded 0.80 for 13 potentially curable classes, excluding negative, *PDE6A*, *PDE6B* and *GUCY2D*; detailed results are presented in Fig. 3b and Supplementary Table 5. The top-5 accuracy for gene classification on internal and external test datasets is 0.904 (95% CI: 0.896–0.912) and 0.856 (95% CI: 0.850–0.863), respectively (Fig. 3c and Supplementary Table 6). We also present the

https://doi.org/10.1038/s41591-026-04545-w

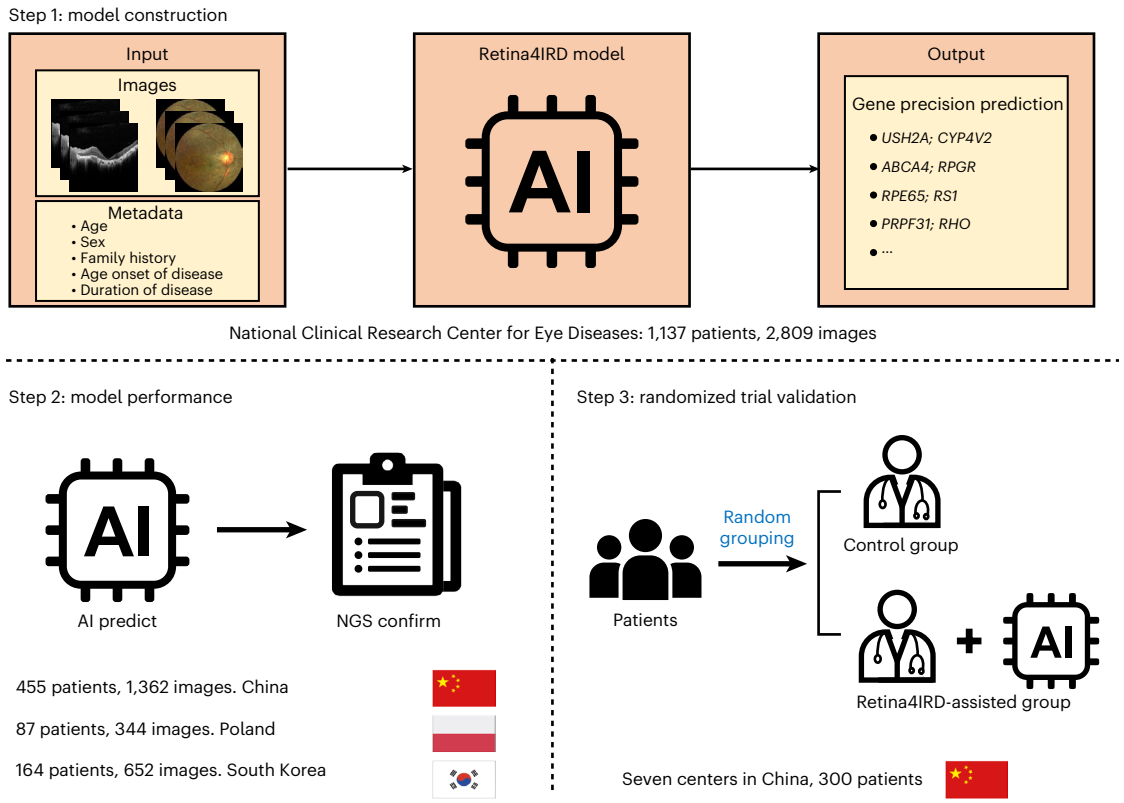


Fig. 1 | Overall study design of the development and validation of the Retina4IRD system. Step 1: study design of the development of the Retina4IRD system. The system accepts fundus images, OCT images and metadata as input and generates predictions for 15 potential treatable gene classes with available clinical trials or approved therapies and untreatable genotypes. Step 2: retrospective multicountry evaluations of the Retina4IRD system in the external validation. In this process, external validation datasets from nine centers across three countries were included. Step 3: a randomized controlled validation trial carried out in China.

site-specific analysis of model performance for the external cohort in Supplementary Fig. 1, and the results were consistent. Progressive integration of metadata with CFP and OCT images showed minimal change in top-5 accuracy, with amplitudes less than 0.01 for both internal (Fig. 3d and Supplementary Table 7) and external (Fig. 3e and Supplementary Table 8) test sets.

Subsequently, ablation studies were conducted to validate the contribution of individual modalities. When combining only CFP with metadata, precision exceeded 0.80 for 10 affected genes, with top-5 accuracies of 0.846 (95% CI: 0.842–0.850) and 0.782 (95% CI: 0.776–0.788) on internal and external test datasets, respectively (Extended Data Fig. 1). While OCT–metadata combination demonstrated enhanced performance, achieving precision exceeded 0.80 for 13 affected genes and perfect accuracy (1.00) for 10 specific genes, containing *ABCA4*, *RPGR*, *RS1*, *PRPF31*, *RHO*, *MERTK*, *CHM*, *CNGA1*, *PDE6A* and *GUCY2D* (Extended Data Fig. 2). This configuration yielded top-5 accuracies of 0.892 (95% CI: 0.881–0.903) internally and 0.868 (95% CI: 0.863–0.872) externally. The detailed results are described in Supplementary Tables 9–13 for CFP single modality and in Supplementary Tables 14–18 for OCT single modality, respectively. Notably, the performance of Retina4IRD based on only images (without clinical metadata) was also outstanding, with top-5 accuracy values of 0.845 (95% CI: 0.841–0.848) and 0.777 (95% CI: 0.768–0.786) for CFP images and 0.889 (95% CI: 0.883–0.895) and 0.866 (95% CI: 0.863–0.870) for OCT images on internal and external datasets, respectively.

Moreover, we calculated the model's prediction performance for the top-3 affected genes (*USH2A*, *CYP4V2* and *ABCA4*) in both the internal and external test sets, with almost all accuracies exceeding 0.85 (Supplementary Fig. 2 and Supplementary Tables 19 and 20). We also validated the performance of Retina4IRD based on international datasets from Poland and South Korea, and the top-5 accuracy remained robust, with values of 0.825 (95% CI: 0.818–0.833) for the Poland cohort and 0.901 (95% CI: 0.895–0.908) for the South Korea cohort (Fig. 3f and Supplementary Table 21). Regarding disease onset and stage, the top-5 accuracy of Retina4IRD in early-onset (age at onset no more than 6 years/12 years), later-onset and end-stage IRD conditions was 0.795/0.821, 0.880 and 0.736, respectively (Supplementary Table 22). Regarding the multi-devices (Supplementary Table 23), Retina4IRD achieved the highest performance on Heidelberg Spectralis ($n = 513$; top-5 accuracy 0.880, 95% CI: 0.873–0.887), followed by ZEISS CIRRUS HD-OCT ($n = 105$; top-5 accuracy 0.819, 95% CI: 0.803–0.835), with lower performance on the 'Other' platforms subgroup ($n = 88$; top-5 accuracy 0.816, 95% CI: 0.808–0.824). Notably, Retina4IRD can also robustly distinguish IRDs from phenotypically similar non-genetic disorders, achieving an area under the curve (AUC) of 0.994 (Supplementary Fig. 3).

Visualization and plausibility of Retina4IRD

The attention maps derived from OCT images highlight the regions most relevant to the model's decision-making, which generally align with pathological areas in clinical experience (Extended Data Fig. 3). Furthermore, deletion and insertion results showed that removing regions highlighted by the model led to a rapid decrease in the prediction score for the ground truth class, whereas inserting these highlighted regions caused the prediction score to increase rapidly, indicating that the model relies on the identified regions for its predictions (Extended Data Fig. 3). In the visualization of model embeddings, uniform manifold approximation and projection (UMAP) of high-dimensional embeddings extracted from OCT scans into two-dimensional space demonstrated overlapping distributions across

https://doi.org/10.1038/s41591-026-04545-w

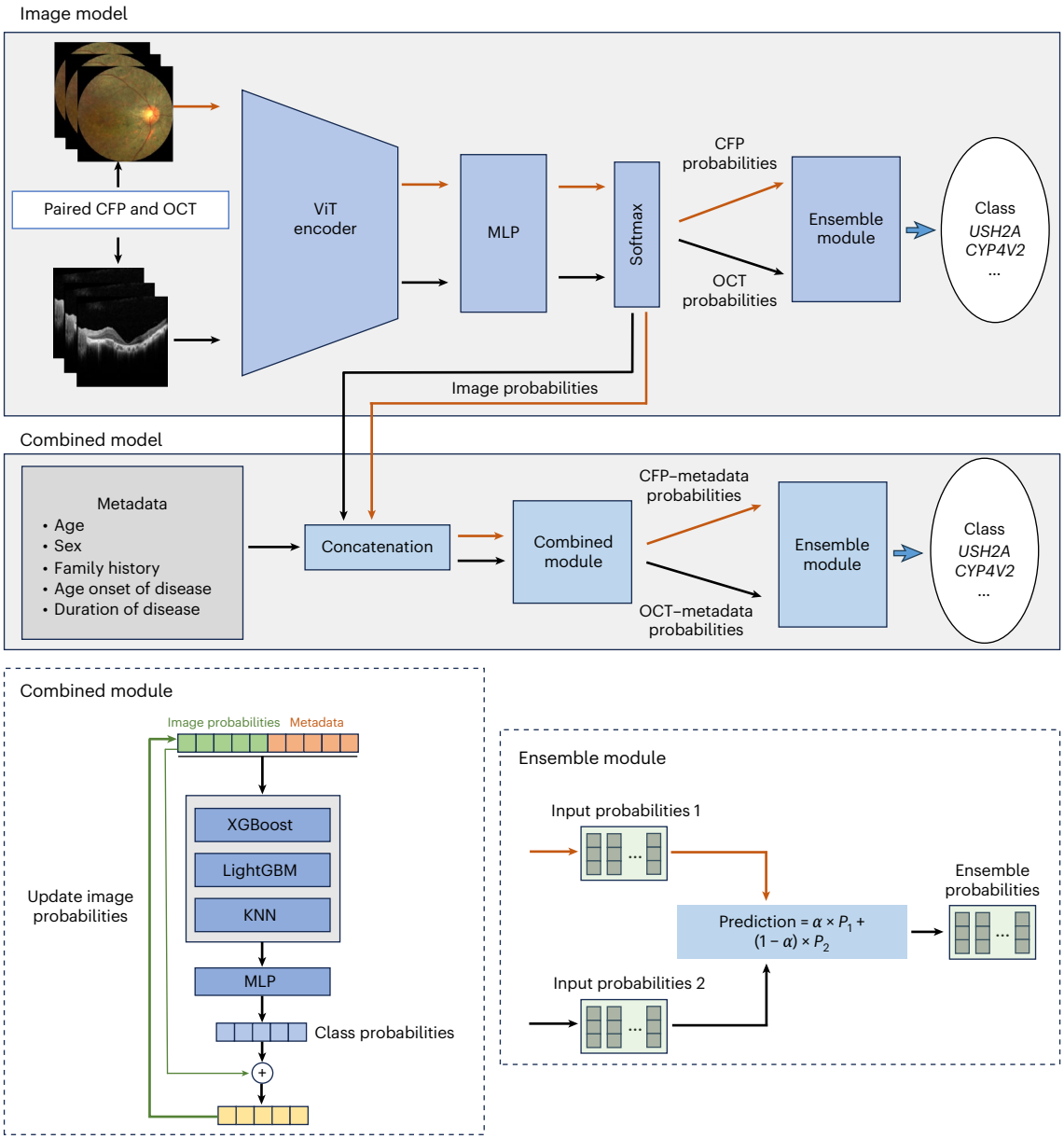


Fig. 2 | Model architecture of the Retina4IRD system. The image model processes CFP and OCT images and produces classwise prediction probabilities. The combined model integrates image-based prediction probabilities with patient metadata through a combined module to generate prediction probabilities. When both CFP and OCT are available, an ensemble module fuses the predictions from the two modalities to produce the final output. ViT, Vision Transformer.

different clinical centers. This indicates the absence of systematic bias in the images (Supplementary Fig. 4).

RCT for the Retina4IRD system

From 27 February to 1 July 2025, a total of 300 participants were enrolled and randomly assigned in a 1:1 allocation ratio using a computer-generated pseudorandom number sequence. Of these, 295 with available NGS reports were included in the final analysis of the RCT, with 148 in the Retina4IRD-assisted specialist arm (AI-assisted group) and 147 in the specialist-only arm (control group). The CONSORT diagram of the trial is presented in Fig. 4. The demographic and clinical characteristics of the patients were well balanced between the two RCT groups (Table 1).

The primary endpoint was defined as top-5 genotype prediction accuracy, which was measured as the proportion of cases in which the correct genotype was ranked among the top-5 predictions. This endpoint was met, with the Retina4IRD-assisted specialist group achieving significantly higher accuracy than the specialist-only control group (88.5% versus 67.3%, $P < 0.001$; Fig. 5). Extended Data Fig. 4 shows that center subgroup analyses revealed consistent superiority of the AI-assisted group, albeit with varying performance patterns: at Shanghai General Hospital, the AI-assisted group maintained a strong advantage (top-5: 91.1% versus 58.7%); Beijing Tongren Hospital showed the largest disparity in top-1 accuracy (46.2% versus 5.6%); and West China Hospital achieved perfect top-5 accuracy in the AI-assisted group (100%).

For the secondary endpoints, the AI-assisted group demonstrated substantially higher accuracy than the control group across all top-1 to top-4 levels, with the most pronounced gap observed at top-1 (37.8% versus 22.4%) and a narrowing gap at top-4 (81.8% versus 53.1%).

Post hoc analyses were performed to further validate the performance of the AI-assisted workflow. Sex subgroup analyses revealed

https://doi.org/10.1038/s41591-026-04545-w

Table 1 | Baseline characteristics of the study population

	Stage 1 (*N*=1,843)						Stage 2 (*N*=295)[a]	
	Training	Validation	Internal test	External test (China)	External test (Poland)	External test (South Korea)	Retina4IRD-assisted group	Control group
Sample size	751	238	148	455	87	164	148	147
Age, years (range)	39 (0, 78)	39 (3, 75)	36 (2, 79)	33 (2, 71)	17 (2, 67)	28 (1, 74)	34 (3, 69)	32 (2, 67)
Sex, *n* (%)								
Males	470 (62.6)	147 (61.8)	88 (59.5)	280 (61.5)	48 (55.2)	94 (57.3)	99 (66.9)	82 (55.8)
Females	281 (37.4)	91 (38.2)	60 (40.5)	175 (38.5)	39 (44.8)	70 (42.7)	49 (33.1)	65 (44.2)
Age onset of disease, years (range)	14 (0, 69)	14 (0, 63)	9 (0, 70)	7 (0, 61)	6 (0, 52)	8 (0, 67)	8 (0, 51)	7 (0, 52)
Duration of disease, years (range)	24 (0, 68)	22 (0, 59)	23 (0, 71)	16 (0, 71)	7 (0, 57)	14 (0, 59)	21 (0, 56)	15 (0, 57)
Family history, *n* (%)								
No	642 (88.4)	208 (88.5)	124 (87.3)	359 (80.0)	65 (74.7)	107 (65.6)	117 (79.1)	120 (81.6)
Yes	84 (11.6)	27 (11.5)	18 (12.7)	90 (20.0)	22 (25.3)	56 (34.4)	31 (20.9)	27 (18.4)
Gene type, *n* (%)								
USH2A	79 (10.5)	27 (11.3)	20 (13.5)	37 (8.1)	9 (10.3)	7 (4.3)	16 (10.8)	12 (8.2)
CYP4V2	48 (6.4)	15 (6.3)	9 (6.1)	21 (4.6)	0 (0.0)	2 (1.2)	7 (4.7)	8 (5.4)
ABCA4	38 (5.1)	12 (5.0)	8 (5.4)	23 (5.1)	9 (10.3)	8 (4.9)	9 (6.1)	9 (6.1)
RPGR	30 (4.0)	9 (3.8)	5 (3.4)	17 (3.7)	11 (12.6)	8 (4.9)	12 (8.1)	4 (2.7)
RPE65	20 (2.7)	8 (3.4)	6 (4.1)	21 (4.6)	0 (0.0)	3 (1.8)	6 (4.1)	4 (2.7)
RS1	21 (2.8)	6 (2.5)	2 (1.4)	15 (3.3)	2 (2.3)	1 (0.6)	3 (2.0)	6 (4.1)
PRPF31	13 (1.7)	4 (1.7)	3 (2.0)	10 (2.2)	2 (2.3)	0 (0.0)	1 (0.7)	0 (0.0)
RHO	14 (1.9)	5 (2.1)	6 (4.1)	13 (2.9)	1 (1.1)	4 (2.4)	5 (3.4)	6 (4.1)
RDH12	10 (1.3)	3 (1.3)	2 (1.4)	9 (2.0)	0 (0.0)	0 (0.0)	1 (0.7)	1 (0.7)
MERTK	11 (1.5)	4 (1.7)	2 (1.4)	7 (1.5)	1 (1.1)	0 (0.0)	2 (1.4)	2 (1.4)
CHM	4 (0.5)	2 (0.8)	3 (2.0)	13 (2.9)	0 (0.0)	6 (3.7)	4 (2.7)	4 (2.7)
CNGA1	10 (1.3)	2 (0.8)	2 (1.4)	10 (2.2)	1 (1.1)	2 (1.2)	4 (2.7)	3 (2.0)
PDE6B	8 (1.1)	3 (1.3)	2 (1.4)	5 (1.1)	3 (3.4)	3 (1.8)	1 (0.7)	2 (1.4)
PDE6A	9 (1.2)	3 (1.3)	1 (0.7)	5 (1.1)	0 (0.0)	0 (0.0)	3 (2.0)	3 (2.0)
GUCY2D	4 (0.5)	2 (0.8)	1 (0.7)	5 (1.1)	0 (0.0)	2 (1.2)	0 (0.0)	1 (0.7)
Incurable	259 (34.5)	80 (33.6)	50 (33.8)	202 (44.4)	48 (55.2)	118 (72.0)	55 (37.2)	61 (41.5)
Negative	173 (23.0)	53 (22.3)	26 (17.6)	42 (9.2)	0 (0.0)	0 (0.0)	19 (12.8)	21 (14.3)

Baseline characteristics are described as median (range) or *n* (%), as appropriate. All results are reported as the number of patients. Family history: 'No' means 'recessive', 'sporadic' or 'unknown'; 'Yes' means 'dominant' or 'X-linked'. [a] In stage 2, a total of 300 participants were enrolled and randomized, of whom 295 with available NGS reports were included in the final analysis.

consistent superiority of the AI-assisted group, with higher top-*N* prediction accuracy compared to the control group in both sexes (Fig. 5). Supplementary Fig. 5 shows that, among participants with a positive genetic report, physicians in the Retina4IRD-assisted arm achieved a higher positive top-5 prediction accuracy than those in the specialist-only arm (86.8% versus 68.3%). Notably, in the AI-assisted group, the workflow consisted of an initial assessment by the specialist followed by a final confirmed result reached with AI assistance. We observed that, without AI assistance, specialists in the AI-assisted group performed similarly to those in the control group in terms of top-5 accuracy (68.2% versus 67.3%), whereas, after incorporating AI assistance, these specialists achieved a significantly higher top-5 accuracy compared to their own unassisted baseline (88.5% versus 68.2%).

In the Retina4IRD-assisted arm, we additionally evaluated the proportion of cases in which clinicians revised their initial diagnostic hypothesis/candidate gene list after reviewing Retina4IRD outputs; under Retina4IRD assistance, the proportion of correct revisions among top-5 gene prediction was 93.6% compared to 13.9% for incorrect revisions, yielding a net decision benefit of 79.7% (Extended Data Fig. 5). Furthermore, post hoc analysis of the composite downstream management score indicated that AI assistance was associated with significantly higher scores relative to the control group (37.7 versus 28.5, $P < 0.001$). The same was true across all four individual dimensions, with mean scores of 10.0 versus 6.6, 9.6 versus 6.8, 8.9 versus 7.8 and 9.2 versus 7.3 for dimensions 1–4, respectively (Extended Data Fig. 5). No adverse events were observed during the trial.

Discussion

The Retina4IRD CDSS system presented in this study represents a potentially transformative advancement in the precision genotype prediction of IRDs. Our approach integrates CFP and OCT, two highly accessible and relatively inexpensive retinal imaging modalities, with clinical metadata through a transformer-based architecture. The model was trained and then validated on a large-scale multicenter dataset comprising 2,138 cases (3,671 eyes) with accompanying whole-exome sequencing (WES) data. This comprehensive evaluation demonstrates the robust capability of the Retina4IRD system to accurately differentiate genetically distinct subtypes of IRDs that currently have therapeutic

https://doi.org/10.1038/s41591-026-04545-w

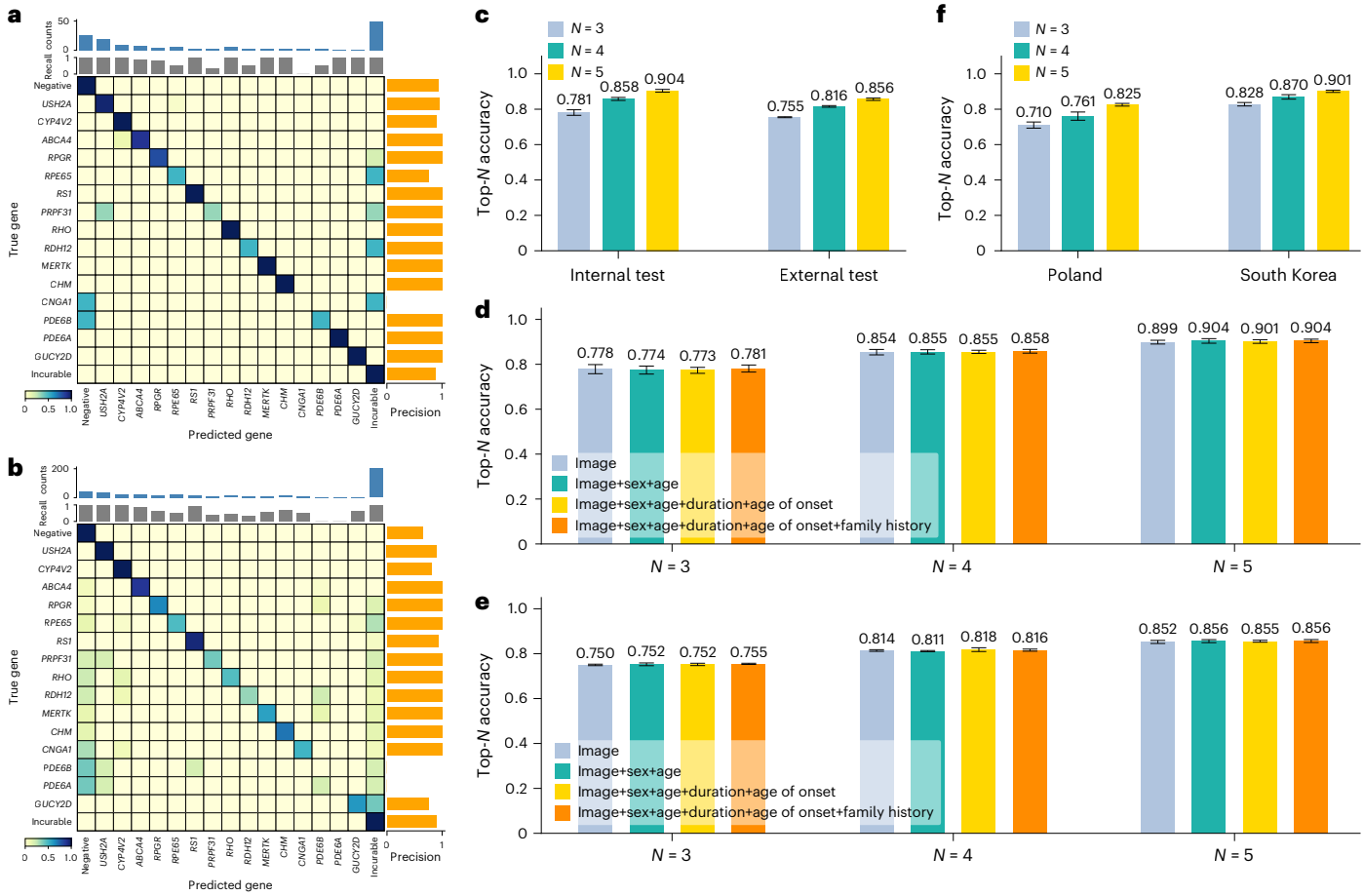


Fig. 3 | Performance of Retina4IRD based on CFP, OCT and metadata. **a**, Evaluation of the model on the internal test set ($n = 148$ cases) across 17 classes using paired CFP and OCT images combined with metadata, including 26 negative cases, 72 potentially curable cases and 50 incurable cases. Patient counts, precision and recall for each class are displayed alongside the confusion matrix. **b**, Evaluation of the external test set ($n = 455$ cases) across 17 classes using paired CFP and OCT images combined with metadata, including 42 negative cases, 211 potentially curable cases and 202 incurable cases. Patient counts, precision and recall for each class are displayed alongside the confusion matrix. **c**, Top-N accuracy ($N \in \{3,4,5\}$) for gene prediction using metadata with CFP and OCT images on the internal test set ($n = 148$ cases) and external test set ($n = 455$ cases). The bar represents the accuracy; the error bar represents 95% CI. **d**, Top-N accuracy ($N \in \{3,4,5\}$) for gene prediction using progressive integration of metadata with CFP and OCT images on the internal test set ($n = 148$ cases). The bar represents the accuracy; the error bar represents 95% CI. **e**, Top-N accuracy ($N \in \{3,4,5\}$) for gene prediction using progressive integration of metadata with CFP and OCT images on the external test set ($n = 455$ cases). The bar represents the accuracy; the error bar represents 95% CI. **f**, Top-N accuracy ($N \in \{3,4,5\}$) for gene prediction on datasets from Poland ($n = 87$ cases) and South Korea ($n = 164$ cases). The bar represents the accuracy; the error bar represents 95% CI.

relevance. Notably, we conducted a prospective RCT and show here that Retina4IRD achieved a remarkable top-5 diagnostic accuracy of 88.5%, significantly surpassing the diagnostic performance of specialist physicians working independently (67.3%). Our results underscore the substantial potential of Retina4IRD as a pre-genetic test CDSS tool that can fit into the clinical workflow of patients with suspected IRD, to assist the physicians' decision to conduct definitive genotype tests of suspected IRDs at both primary eyecare and even specialist-level settings.

The diagnostic challenges inherent to IRDs exemplify what might be termed 'the universal predicament of rare diseases', characterized by limited patient cohorts and insufficient familiarity that most clinicians have with IRDs. Even for highly experienced specialists, the accurate diagnosis of atypical cases frequently necessitates reliance on comprehensive genetic testing[21–25]. Regrettably, over one-third of patients with IRD worldwide fail to obtain definitive diagnoses due to prohibitive costs and protracted diagnostic timelines, resulting in missed opportunities for gene therapy or clinical trial enrollment[7,26]. The Retina4IRD addresses this critical unmet need through an innovative AI-powered prediction and targeted validation paradigm. By generating a probability-ranked list of candidate pathogenic genes, the system enables clinicians or patients to selectively validate only the top-N high-probability targets. This approach substantially reduces unnecessary testing while lowering healthcare costs and accelerating precision treatment. Notably, Retina4IRD provides essential visualization and plausibility for decision-making through attention heatmaps, deletion and insertion experiments and embedding space visualization[27–29], which mitigates application risks more effectively while building clinical trust.

The current landscape of AI applications to assist genotype diagnosis for patients with IRD remains remarkably sparse; most studies were limited by modest sample sizes and restricted genetic coverage[16–20,30]. Among these, the Eye2Gene model is a representative AI system for IRD gene prediction to date. However, although Eye2Gene was trained predominantly on non-Caucasian populations, it included only 0.18% Chinese patient data. By contrast, our Retina4IRD system was developed exclusively using Chinese datasets, representing one-fifth of the global population, thereby addressing a critical gap in ethnic representation. Additionally, Retina4IRD represents several key advances. First, it is built upon the Vision Transformer architecture[31,32] and leverages pretrained weights from the visual foundation model RETFound, which was pretrained on over 0.9 million CFP

https://doi.org/10.1038/s41591-026-04545-w

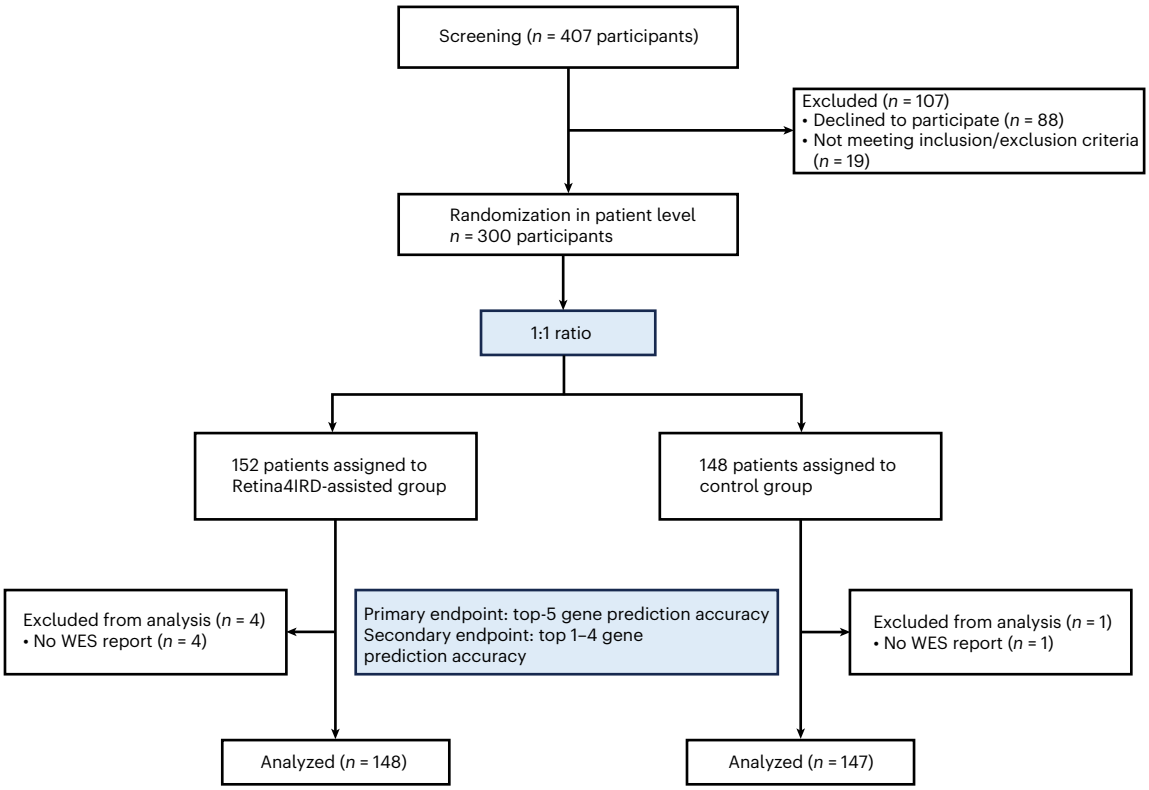


Fig. 4 | CONSORT diagram for the RCT. In total, 407 individuals were screened, and a total of 300 participants were enrolled and randomly assigned in a 1:1 allocation ratio using a computer-generated pseudorandom number sequence. Of these, 295 with available NGS reports were included in the final analysis of the RCT, with 148 in the Retina4IRD-assisted specialist arm (AI-assisted group) and 147 in the specialist-only arm (control group). The primary outcome measure was top-5 gene prediction accuracy, defined as the proportion of cases in which the correct genotype was ranked among the top-5 predictions. Specifically, a ranked list of candidate genotypes was generated for each case by the two arms. A prediction was considered correct if the actual genotype, as confirmed by the subsequent NGS report (the gold standard), fell within the top-5 ranked predictions. Secondary outcomes included the accuracy of top-1 to top-4 gene predictions, defined as the proportion of cases in which the correct genotype prediction was ranked among the top-*N* predictions. A prediction was considered correct if the actual genotype, as confirmed by the subsequent NGS report, fell within the top-*N* ranked predictions.

and 0.7 million OCT images. Compared to a combination attention–convolutional neural network (CNN) model (Eye2Gene), Retina4IRD may help capture clinically relevant features from both CFP and OCT images, which plays a critical role in improving the accuracy of genotype prediction[33]. Unlike image-only models of Eye2Gene, Retina4IRD integrates multimodal data, including CFP, OCT and clinical metadata, into a unified framework. This multimodal fusion more closely reflects the diagnostic reasoning process of clinical experts when evaluating IRDs, leading to improved model reliability and clinical relevance. Second, although Retina4IRD predicts fewer genes than Eye2Gene, it strategically prioritizes 15 potentially treatable pathogenic variants, referring to those with clinical trials or approved therapy, responsible for approximately 60% of IRD cases. This target diagnostic approach demonstrates greater clinical utility, as confirming genotypes without effective treatments via costly NGS lacks urgency, particularly in resource-limited settings. Lastly, external validation underscores the exceptional precision of Retina4IRD, with prediction accuracy exceeding 85% for key therapeutic target genes, including *USH2A*, *ABCA4*, *RPGR*, *RPE65*, *RS1*, *PRPF31*, *RHO*, *RDH12*, *MERTK*, *CHM* and *CNGA1*, and maintaining strong concordance with NGS results. The robust generalizability was further confirmed through cross-ethnic validation in both Poland and South Korea cohorts.

Notably, Retina4IRD represents, to our knowledge, the first and only AI system capable of predicting multiple target genes using CFP—the most accessible and cost-effective retinal imaging modality. This represents a potentially major advancement over previous AI models for IRD genotype prediction, which relied exclusively on fundus autofluorescence (FAF) or OCT imaging[16–20,30]. This technological breakthrough carries important clinical implications. FAF imaging systems remain prohibitively expensive and are typically available only at tertiary referral centers. By contrast, Retina4IRD overcomes these equipment barriers by leveraging widely available CFP/OCT devices, which are already standard in most ophthalmic clinics. It achieves a top-5 gene prediction accuracy of 90.4% in internal test dataset, whereas Eye2Gene reported a top-5 accuracy of 83.9% in its own study[20]. These findings align with growing evidence supporting the diagnostic utility of routine ocular images, including recent work highlighting their role in distinguishing between rare and common fundus diseases[34]. Notably, our system extends this capability by achieving exceptional performance in distinguishing IRDs from phenotypically similar non-genetic disorders using only CFP imaging. This enables practical implementation in primary care settings or resource-constrained regions with specialist availability.

To establish definitive clinical validity, we conducted a prospective multicenter RCT evaluating our AI system, Retina4IRD, for genotype prediction of IRD in authentic clinical practice settings. Our study presents one of the first prospective RCTs to test an AI system within actual clinical workflows. This step is crucial to properly evaluate their real-world performance and limitations[35,36]. Unlike previous retrospective investigations, our trial was designed to reflect real-world clinical practice, providing robust evidence of the model's clinical value. The results demonstrated the superior capability of Retina4IRD in assisting with genotype diagnosis, achieving 88.5% top-5 accuracy in pathogenic gene prediction, significantly outperforming the 67.3%

https://doi.org/10.1038/s41591-026-04545-w

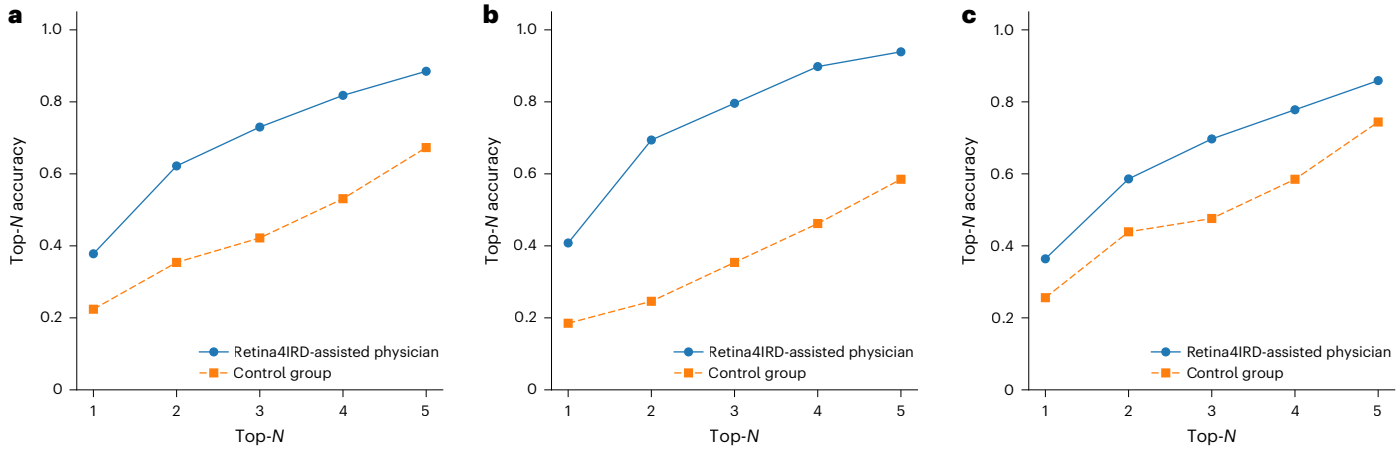


Fig. 5 | Top-*N* accuracy in FAS and sex subgroup. a–**c**, Accuracy for the overall cohort ($n = 295$; **a**), female subgroup ($n = 114$; **b**) and male subgroup ($n = 181$; **c**), respectively. The sample sizes for the Retina4IRD-assisted group and the control group were 148 versus 147 for **a**, 49 versus 65 for **b** and 99 versus 82 for **c**, respectively. Points represent the accuracy. The *P* values were calculated via χ^2 test using the top-5 accuracy levels, with the values of <0.001, <0.001 and 0.079 for **a**, **b** and **c**, respectively. $N \in \{1,2,3,4,5\}$.

accuracy attained by clinician only. Notably, too, this trial was conducted at tertiary referral centers staffed by retinal specialists with extensive IRD expertise, yet it still revealed the considerable challenge that genotype–phenotype correlation presents even for highly experienced clinicians during routine patient care. Our study provides, to our knowledge, the first prospective comparison between AI and genetic retinal specialists for genotype prediction in IRD cases and delivers evidence meeting the gold standard for clinical validation. The successful translation from algorithm development to clinical application marks a critical advancement, moving beyond proof-of-concept research to establish Retina4IRD as a practical tool that genuinely enhances both accuracy and efficiency in ophthalmic practice. The rigorous methodology and noteworthy outcomes of this trial set a new benchmark for validating AI systems in specialized medical fields, particularly for complex genetic disorders such as IRDs.

Although demonstrating promising performance and clinical utility, several limitations of the current Retina4IRD system warrant consideration. First, it must be emphasized that the predictions of Retina4IRD serve as a pre-genetic test CDSS tool rather than a replacement for definitive genetic testing and expert genetic counseling, particularly when determining eligibility for gene therapy or other treatment interventions requiring definitive molecular diagnosis. The current version of Retina4IRD focuses on 17 genotype categories selected for established therapeutic relevance. Although these account for approximately 60% of IRD cases in our validation cohort, future work should expand coverage to additional rare genes using more diverse datasets, as novel therapies emerge for more genetic subtypes. Notably, some genes exhibit both dominant and recessive inheritance patterns, which can limit the usefulness of simplified clinical metadata. For example, family history is captured as a binary variable in our model and should be interpreted with caution: a 'yes' response may suggest dominant or X-linked inheritance, whereas a 'no' response may still include recessive, sporadic or de novo dominant cases. Second, methodologically, Retina4IRD prioritizes robustness through a late-fusion design that tolerates missing modalities and relies primarily on the central OCT B-scan; these strategies may limit its ability to fully capture crossmodal feature interactions. Exploring feature-level fusion or joint multimodal representation learning is an important direction for future work. Moreover, the entropy-weighted fusion method relies on the assumption that lower predictive entropy reflects higher diagnostic reliability; however, this assumption may be violated under miscalibration or domain shift. In addition, although Retina4IRD was evaluated on multiple devices, performance may vary by scanner, particularly for underrepresented platforms, so results for less represented or unseen devices should be interpreted cautiously. Third, our Retina4IRD model was trained primarily on a relatively homogeneous population from China, and our multicenter RCT was conducted exclusively in Chinese patients. Although Retina4IRD maintained approximately 80% top-5 accuracy in the external validation cohorts from South Korea and Poland, such evaluations cannot fully substitute for formal generalizability evidence obtained across different healthcare systems and population backgrounds. Finally, we clarify that our composite downstream management score is an exploratory post hoc analysis, reflecting the management plan proposed at the clinic visit. It does not directly demonstrate improvements in long-term clinical outcomes. Therefore, multinational prospective trials are warranted to confirm the generalizability of the clinical impact across broader ethnic and geographic populations and care settings and to assess the long-term clinical benefits of these management strategies, informing future development of Retina4IRD toward a more comprehensive disease management platform.

In conclusion, our study establishes the potential of Retina4IRD as a useful AI-assisted pre-genetic test CDSS tool for precision genotyping of IRD, improving the decision-making processes of physicians at primary care and specialist levels. Our system achieves accurate, interpretable and generalizable performance, effectively transforming the traditional time-consuming and resource-intensive diagnostic pathway into a more cost-effective and efficient clinical workflow.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-026-04545-w.

References

1. Georgiou, M. et al. Phenotyping and genotyping inherited retinal diseases: molecular genetics, clinical and imaging features, and therapeutics of macular dystrophies, cone and cone-rod dystrophies, rod-cone dystrophies, Leber congenital amaurosis, and cone dysfunction syndromes. *Prog. Retin. Eye Res.* **100**, 101244 (2024).
2. Ben-Yosef, T. Inherited retinal diseases. *Int. J. Mol. Sci.* **23**, 13467 (2022).

3. Schneider, N. et al. Inherited retinal diseases: linking genes, disease-causing variants, and relevant therapeutic modalities. *Prog. Retin. Eye Res.* **89**, 101029 (2022).
4. Sullivan, L. S. & Daiger, S. P. RetNet: Retinal Information Network. https://retnet.org/
5. Britten-Jones, A. C. et al. The diagnostic yield of next generation sequencing in inherited retinal diseases: a systematic review and meta-analysis. *Am. J. Ophthalmol.* **249**, 57–73 (2023).
6. Sheck, L. H. N. et al. Panel-based genetic testing for inherited retinal disease screening 176 genes. *Mol. Genet. Genomic Med.* **9**, e1663 (2021).
7. Nguyen, Q. et al. Can artificial intelligence accelerate the diagnosis of inherited retinal diseases? Protocol for a data-only retrospective cohort study (Eye2Gene). *BMJ Open* **13**, e071043 (2023).
8. Wong, W. et al. Inherited retinal disease pathway in the UK: a patient perspective and the potential of AI. *Br. J. Ophthalmol.* **109**, 1266–1271 (2025).
9. Pierce, E. A. et al. Gene editing for *CEP290*-associated retinal degeneration. *N. Engl. J. Med.* **390**, 1972–1984 (2024).
10. Botto, C. et al. Early and late stage gene therapy interventions for inherited retinal degenerations. *Prog. Retin. Eye Res.* **86**, 100975 (2022).
11. Suh, S., Choi, E. H., Raguram, A., Liu, D. R. & Palczewski, K. Precision genome editing in the eye. *Proc. Natl Acad. Sci. USA* **119**, e2210104119 (2022).
12. 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).
13. Ting, D. S. W. et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA* **318**, 2211–2223 (2017).
14. Dai, L. et al. A deep learning system for predicting time to progression of diabetic retinopathy. *Nat. Med.* **30**, 584–594 (2024).
15. Meng, Z. et al. Non-invasive biopsy diagnosis of diabetic kidney disease via deep learning applied to retinal images: a population-based study. *Lancet Digit. Health* **7**, 100868 (2025).
16. Fujinami-Yokokawa, Y. et al. Prediction of causative genes in inherited retinal disorder from fundus photography and autofluorescence imaging using deep learning techniques. *Br. J. Ophthalmol.* **105**, 1272–1279 (2021).
17. Fujinami-Yokokawa, Y. et al. Prediction of causative genes in inherited retinal disorders from spectral-domain optical coherence tomography utilizing deep learning techniques. *J. Ophthalmol.* **2019**, 1691064 (2019).
18. Shah, M., Roomans Ledo, A. & Rittscher, J. Automated classification of normal and Stargardt disease optical coherence tomography images using deep learning. *Acta Ophthalmol.* **98**, e715–e721 (2020).
19. Miere, A. et al. Deep learning to distinguish *ABCA4*-related Stargardt disease from *PRPH2*-related pseudo-Stargardt pattern dystrophy. *J. Clin. Med.* **10**, 5742 (2021).
20. Pontikos, N. et al. Next-generation phenotyping of inherited retinal diseases from multimodal imaging with Eye2Gene. *Nat. Mach. Intell.* **7**, 967–978 (2025).
21. 100,000 Genomes Project Pilot Investigators et al. 100,000 Genomes pilot on rare-disease diagnosis in health care—preliminary report. *N. Engl. J. Med.* **385**, 1868–1880 (2021).
22. Conway, M. P. et al. The role of the ophthalmic genetics multidisciplinary team in the management of inherited retinal degenerations—a case-based review. *Life (Basel)* **14**, 107 (2024).
23. Parker, M. A. et al. Test–retest variability of functional and structural parameters in patients with Stargardt disease participating in the SAR422459 gene therapy trial. *Transl. Vis. Sci. Technol.* **5**, 10 (2016).
24. Guziewicz, K. E. et al. *BEST1* gene therapy corrects a diffuse retina-wide microdetachment modulated by light exposure. *Proc. Natl Acad. Sci. USA* **115**, E2839–E2848 (2018).
25. Rahman, N., Georgiou, M., Khan, K. N. & Michaelides, M. Macular dystrophies: clinical and imaging features, molecular genetics and therapeutic options. *Br. J. Ophthalmol.* **104**, 451–460 (2020).
26. A preliminary assessment of the potential impact of rare diseases on the NHS. https://imperialcollegehealthpartners.com/a-preliminary-assessment-of-the-potential-impact-of-rare-diseases-on-the-nhs/ (Imperial College Health Partners, 2018).
27. Reddy, S. Explainability and artificial intelligence in medicine. *Lancet Digit. Health* **4**, e214–e215 (2022).
28. Kundu, S. AI in medicine must be explainable. *Nat. Med.* **27**, 1328 (2021).
29. Desai, A. N. Artificial intelligence: promise, pitfalls, and perspective. *JAMA* **323**, 2448–2449 (2020).
30. Miere, A. et al. Deep learning-based classification of inherited retinal diseases using fundus autofluorescence. *J. Clin. Med.* **9**, 3303 (2020).
31. Zhou, Y. et al. A foundation model for generalizable disease detection from retinal images. *Nature* **622**, 156–163 (2023).
32. Kim, C. et al. Transparent medical image AI via an image-text foundation model grounded in medical literature. *Nat. Med.* **30**, 1154–1165 (2024).
33. Guo, L. L. et al. A multi-center study on the adaptability of a shared foundation model for electronic health records. *NPJ Digit. Med.* **7**, 171 (2024).
34. Wang, M. et al. Enhancing diagnostic accuracy in rare and common fundus diseases with a knowledge-rich vision-language model. *Nat. Commun.* **16**, 5528 (2025).
35. Han, R. et al. Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. *Lancet Digit. Health* **6**, e367–e373 (2024).
36. Angus, D. C. Randomized clinical trials of artificial intelligence. *JAMA* **323**, 1043–1045 (2020).

[1]Department of Ophthalmology, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai Key Laboratory of Fundus Diseases, National Clinical Research Center for Eye Diseases, Shanghai, China. [2]College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, Key Laboratory of Brain-Machine Intelligence Technology, Ministry of Education, Nanjing, China. [3]Beijing Visual Science and Translational Eye Research Institute (BERI), Beijing Tsinghua Changgung Hospital Eye Center, School of Clinical Medicine, Tsinghua Medicine, Tsinghua University, Beijing, China. [4]Institute of Vision Research, Department of Ophthalmology, Severance Hospital, Yonsei University College of Medicine, Seoul, Republic of Korea. [5]Shanghai International Joint Laboratory for Intelligent Prevention and Treatment of Metabolic Disorders, Department of Computer Science

and Engineering, MOE Key Laboratory of AI, School of Electronic, Information, and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China. [6]Institute of Ophthalmology, Beijing Tongren Eye Center, Beijing Tongren Hospital, Capital Medical University, Beijing Ophthalmology & Visual Sciences Key Lab, Beijing, China. [7]Department of Ophthalmology, Eye and ENT Hospital, Fudan University, Shanghai Key Laboratory of Visual Impairment and Restoration, Science and Technology Commission of Shanghai Municipality, Key Laboratory of Myopia and Related Eye Diseases, NHC, Chinese Academy of Medical Sciences, Shanghai, China. [8]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science, Guangzhou, China. [9]Department of Ophthalmology, West China Hospital, Sichuan University, Chengdu, China. [10]Department of Ophthalmology, Ningxia Eye Hospital, People's Hospital of Ningxia Hui Autonomous Region, Yinchuan, China. [11]Henan Eye Hospital, Henan Provincial People's Hospital, Zhengzhou, China. [12]Eye Institute, Henan Academy of Innovations in Medical Science, Zhengzhou, China. [13]Institute of Vision Research, Department of Ophthalmology, Gangnam Severance Hospital, Yonsei University College of Medicine, Seoul, Republic of Korea. [14]Department of Pediatric Ophthalmology, Faculty of Medical Sciences in Katowice, Medical University of Silesia, Katowice, Poland. [15]Department of Pediatric Ophthalmology, Kornel Gibinski University Clinical Center, Medical University of Silesia, Katowice, Poland. [16]GlaucoTech Co., Katowice, Poland. [17]Department of Clinical Genetics and Rare Diseases, Faculty of Medical Sciences in Katowice, Medical University of Silesia, Katowice, Poland. [18]Department of Ophthalmology, Kornel Gibinski University Clinical Center, Medical University of Silesia, Katowice, Poland. [19]Ophthalmology and Visual Sciences Academic Clinical Program (Eye ACP), Duke-NUS Medical School, Singapore, Singapore. [20]Medi Whale Inc. Seoul, Seoul, Republic of Korea. [21]Department of Ophthalmology, University of Warmia and Mazury, Olsztyn, Poland. [22]Institute for Research in Ophthalmology, Foundation for Ophthalmology Development, Poznan, Poland. [23]School of Biomedical Engineering, Tsinghua Medicine, Tsinghua University, Beijing, China. [24]Singapore Eye Research Institute, Singapore National Eye Center, Singapore, Singapore. [25]These authors contributed equally: Huixun Jia, Bo Qian, Yanlin Qu, Jinyuan Wang, Jieqiong Chen, Tong Li, Changle Zheng, Jinu Han. ✉e-mail: shengbin@sjtu.edu.cn; wongtienyin@tsinghua.edu.cn; xdsun@sjtu.edu.cn

Methods

Ethical approval

This study was conducted in accordance with all applicable ethical regulations. Stage 1 was approved by the Ethics Committee of the National Clinical Research Center for Eye Diseases (2022KY140, approved 30 December 2022; 2022KY140-2, approved 29 December 2023). As only deidentified retrospective data were used, with no active participant involvement, the need for informed consent was waived for these data and images. Stage 2 was approved by the same Ethics Committee (2024-451, approved 6 November 2024) and was registered with ClinicalTrials.gov (NCT06839170). Written informed consent was obtained from all participants. All procedures were performed in compliance with the Declaration of Helsinki. The detailed protocol is shown in Supplementary Information.

Study design and datasets

This is a two-stage study (Fig. 1). Stage 1 is model development and evaluation, involving 5,167 images and 8,901 metadata items from 1,843 genetically confirmed individuals (median age 36 years, 716 females, 3,376 eyes) across China, South Korea and Poland. Stage 2 is a randomized controlled validation trial involving 295 participants (median age 33 years, 114 females) conducted to evaluate the effectiveness of Retina4IRD in a real-world clinical setting. This study strictly adheres to Sex and Gender Equity in Research (SAGER) guidelines. Consistent with these recommendations, we have applied the term 'sex' solely in reference to biological attributes. Participant sex was determined from official identity documentation, and the data were classified into two categories: female and male.

Stage 1: model development and evaluation

In stage 1, the Retina4IRD was initially constructed via a multimodal dataset comprising 2,809 images (CFP and OCT scans) along with 5,538 clinical metadata items (age, sex, family history, age of onset and duration of disease). The data were derived from 1,137 genetically confirmed cases (2,197 eyes) collected at the National Clinical Research Center for Eye Diseases. The dataset was randomly split as follows: training set, 66% of cases (n = 751, 1,461 eyes); internal validation set, 21% of cases (n = 238, 453 eyes); and internal test set, 13% of cases (n = 148, 283 eyes). External validation was performed on 455 participants (681 eyes) from seven tertiary ophthalmic centers across China: Eye & ENT Hospital of Fudan University; Beijing Tongren Hospital; Zhongshan Ophthalmic Center, Sun Yat-sen University; Shanghai General Hospital; West China Hospital, Sichuan University; Henan Provincial People's Hospital; and People's Hospital of Ningxia Hui Autonomous Region, comprising 1,362 images and 2,239 metadata items. To further assess cross-population generalizability, additional validation was conducted using international datasets: 164 genetically confirmed cases (326 eyes) with 652 images and 817 metadata items from Severance Hospital at Yonsei University in South Korea and 87 genetically confirmed cases (172 eyes) with 344 images and 307 metadata items from the University of Warmia and Mazury in Poland. We used fundus images acquired across participating centers on multiple commonly used platforms, without restricting the model to a single manufacturer or device (for example, CFP cameras such as Canon CR-2/CR-2 Plus, Topcon TRC-50DX or ZEISS VISUCAM and OCT systems such as Heidelberg Spectralis, ZEISS CIRRUS HD-OCT or Topcon DRI OCT Triton). There was no overlap among patients in the training dataset, the internal validation dataset or the external validation datasets.

Model development

The overall architecture of Retina4IRD is illustrated in Fig. 2. The framework consists of two components: an image model and a combined model. The image models are used to generate prediction results when CFP and OCT images are provided as inputs. When metadata are also available, the combined model integrates the outputs of the image model with the metadata to produce the final predictions. The combined model is trained separately from the image model, after the image model training is completed.

Image model. For the CFP and OCT modalities, separate models are independently trained using the same network architecture. Each input image is first divided into nonoverlapping 16 × 16 patches, and each patch is projected into a corresponding embedding vector through a linear projection layer. These embedding vectors serve as inputs to the encoder and are used to extract high-level visual representations for each patch. The encoder adopts the large version of Vision Transformer, which consists of 24 transformer blocks. The resulting visual features are then aggregated via global average pooling to form a unified feature representation, which is subsequently fed into a multilayer perceptron (MLP) to produce classwise prediction probabilities. For each imaging modality, if a patient has images from both eyes, an entropy-weighted fusion strategy is applied to combine the eye-level predictions into a patient-level prediction. Through this process, both CFP and OCT models generate patient-level classwise probability distributions. Finally, a model ensemble approach is applied to integrate predictions from the two modalities, where the final class probability distribution is obtained by averaging the patient-level probabilities from the CFP and OCT models.

During the training phase, a variety of data augmentation techniques are employed to enhance the model's robustness and generalization ability. These include brightness adjustment, contrast enhancement, rotation, scaling and random erasing. The augmented images are then normalized and resized to a resolution of 224 × 224 pixels before being fed into the network. During the testing phase, input images are center-cropped to 224 × 224 and normalized using the same parameters applied during training to ensure consistency. The model is independently trained using five different random seeds, each for 50 epochs with a batch size of 32. The first 10 epochs involve learning rate warming up from 0 to 5×10^{-4}, followed by a cosine annealing schedule reducing the learning rate from 5×10^{-4} to 1×10^{-6} over the remaining 40 epochs. After each epoch, the model is evaluated on the validation set, and the weights with the best performance are saved for subsequent evaluation on the test set.

Combined model. The metadata are concatenated with the patient-level classwise prediction probabilities generated by the image model and used as input to a combined module consisting of three classifiers: XGBoost, LightGBM and k-nearest neighbors (KNN). XGBoost employs the Dropouts meet Additive Regression Trees (DART) algorithm together with L2 regularization to improve generalization performance. LightGBM is configured with a predefined maximum number of iterations and incorporates L1 regularization to control model complexity and mitigate overfitting. For KNN, the number of neighbors is set to 9. The outputs of the three classifiers are subsequently fed into an MLP to generate classwise prediction probabilities.

To further improve classification performance, the combined module adopts an iterative refinement training strategy. Specifically, the classwise image probabilities produced by the image model for a given modality are denoted by P_I. These probabilities are concatenated with the metadata to form the combined input, which is then processed by the classifiers (XGBoost, LightGBM and KNN) and subsequently passed through the MLP to generate class probabilities P_{MLP}. The fused output is computed as

$$P_{\mathrm{output}} = \beta \times P_I + (1 - \beta) \times P_{\mathrm{MLP}}$$

The fused probability P_{output} then replaces P_I and is used as the input probability for the next iteration, and the above procedure is repeated multiple times. Grid search is used to determine the optimal fusion weight β and the number of refinement iterations. In the final

model, the optimal value of β is 0.974, and the refinement process is performed for six training cycles.

For each imaging modality, if a patient has images from both eyes, an entropy-weighted fusion strategy is applied to combine the eye-level predictions into a patient-level prediction. If a patient has both metadata and images from the two modalities, the combined model is applied to each image modality separately to generate patient-level prediction probabilities. The probabilities obtained from the two modalities are then integrated using an ensemble strategy based on probability averaging, to obtain the final gene probability distribution.

XGBoost is configured with 21 boosting estimators and incorporates L2 regularization by setting the λ parameter to 0.46. For LightGBM, the number of boosting iterations is set to 33, with a maximum of eight leaves per decision tree, and the corresponding L1 regularization parameter is set to 0.21. We employ an MLP with a single hidden layer consisting of 100 neurons. The initial learning rate is set to 0.0186 to facilitate efficient convergence during optimization. Prior to training, all continuous features are standardized using StandardScaler. The validation and test sets use the same scaling parameters as the training set to ensure consistent input distributions across all datasets.

Entropy-weighted fusion strategy. For patients with images from both eyes, we combined the predictions from the two eyes to obtain a patient-level genotype prediction using an entropy-weighted fusion strategy. This approach assigns higher weights to predictions with lower predictive uncertainty.

The entropy-weighted fusion strategy allows the model to integrate information from both eyes while automatically assigning greater influence to the eye with higher predictive confidence, leading to more reliable patient-level predictions. The specific processing details are as follows:

Let K denote the number of candidate genotypes. For each eye, the model outputs a probability distribution over the K genotypes:

$$\mathbf{p}^L = (p_1^L, p_2^L, \ldots, p_K^L), \mathbf{p}^R = (p_1^R, p_2^R, \ldots, p_K^R)$$

where $\sum_{k=1}^{K} p_k^L = 1$ and $\sum_{k=1}^{K} p_k^R = 1$.

We first compute the predictive entropy for each eye:

$$H_L = -\sum_{k=1}^{K} p_k^L \ln p_k^L, H_R = -\sum_{k=1}^{K} p_k^R \ln p_k^R$$

Lower entropy indicates a more confident prediction. To quantify the reliability of each eye's prediction, we define a quality score based on normalized entropy:

$$q_L = 1 - \frac{H_L}{\ln K}, q_R = 1 - \frac{H_R}{\ln K}$$

Finally, the patient-level genotype probability is obtained by combining the two eye-level predictions in the log-probability space:

$$s_k = w_L \ln p_k^L + w_R \ln p_k^R$$

$$p_k^P = \frac{e^{s_k}}{\sum_{j=1}^{K} e^{s_j}}$$

where $\mathbf{p}^P = (p_1^P, p_2^P, \ldots, p_K^P)$ represents the patient-level probability distribution over the K candidate genotypes.

This entropy-weighted fusion strategy allows the model to integrate information from both eyes while automatically assigning greater influence to the eye with higher predictive confidence, leading to more reliable patient-level predictions.

For clinical metadata with missing values, statistical imputation methods were applied. Missing values for continuous variables, disease duration and age at onset were filled with the mean value calculated from patients with the same genotype. After imputation, logical consistency checks were performed to ensure clinical plausibility: if the imputed age of onset exceeded the patient's current age, it was replaced with the current age. For family history, missing values were imputed using the most frequent category among patients with the same genotype.

Visualization and plausibility of Retina4IRD

To address the limited visualization of the AI model, we leveraged the self-attention mechanism of the Retina4IRD architecture to generate attention maps from OCT images. These maps highlight image regions that receive the highest attention weights and thus contribute most to the model's predictions. We further conducted deletion and insertion experiments to assess the faithfulness of the attention maps. In the deletion experiment, pixels with the highest attention scores were progressively removed from the original image, gradually eliminating the regions that most strongly influence the model's predictions. In the insertion experiment, we started from a blank image and progressively restored pixels with the highest attention scores. At each step, the modified image was fed into the model, and the model's confidence for the ground truth class was recorded.

For assessing class diversity and revealing interclass similarities, we employed a data-driven strategy using the Retina4IRD model to process all OCT scans from both internal and external test datasets. From the penultimate layer activations, we extracted 768-dimensional feature vectors and projected them into two-dimensional space using UMAP for visualization[37]. Overlapping distributions of data across different clinical centers in the UMAP visualization indicate that the model has minimal systematic bias.

Stage 2: randomized controlled validation trial

For stage 2, a randomized controlled validation trial was conducted to evaluate the performance of Retina4IRD (also designated as FM-IRD) in clinical settings. This trial was conducted at seven centers in China with large outpatient IRD care (Shanghai General Hospital; Eye & ENT Hospital of Fudan University; Beijing Tongren Hospital; Zhongshan Ophthalmic Center, Sun Yat-sen University; West China Hospital, Sichuan University; Henan Provincial People's Hospital; and People's Hospital of Ningxia Hui Autonomous Region), and the details are presented in Supplementary Information. Participants were recruited from the outpatient of IRD clinics of the participating centers or through referrals from collaborating retinal specialists. In addition, we actively recruited referred patients with suspected IRD from external centers, predominantly referred by retina specialists with a high pre-test suspicion of IRD.

All participants underwent NGS-based testing, specifically WES. The report turnaround time is defined as the interval from sample submission to issuance of the final genetic report. The genetic testing workflow included sample submission, DNA library preparation (2–3 days), sequencing (2 days), variant interpretation (2–5 days) and family segregation analysis (3–5 days), where indicated, and report issuance (1–2 days). The reporting timeline was required not to exceed 30 days under normal circumstances. Given that genetic testing results were unavailable at enrollment, and to prioritize the ability of the model to identify patients with actionable therapeutic targets, genotypes of our RCT were categorized into 17 classes, including a negative class, comprising participants for whom NGS did not identify a causative gene; genotypes with potentially available gene therapies or clinical trials (Supplementary Table 24, including *USH2A*, *CYP4V2*, *ABCA4*, *RPGR*, *RPE65*, *RS1*, *PRPF31*, *RHO*, *RDH12*, *MERTK*, *CHM*, *CNGA1*, *PDE6B*, *PDE6A* and *GUCY2D*); and currently untreatable genotypes. All participants provided written informed consent before enrollment assessments.

The inclusion criterion was patients clinically suspected of having IRD during outpatient retina clinic visits. Exclusion criteria included refusal to undergo NGS genetic testing, fundus photography or OCT

examination; presence of other ocular conditions that may interfere with retinal examination; history of intraocular surgery in either eye within 6 months prior to screening; systemic severe diseases, intellectual disability or psychiatric disorders; and participation in other interventional clinical trials, such as stem cell or gene therapy trials. The study was reviewed and approved by the appropriate ethics review boards and was registered at ClinicalTrials.gov (NCT06839170).

Randomization and masking. Randomization was performed in a 1:1 ratio using a computer-generated pseudorandom number sequence by an independent statistician. After eligibility assessment, all participants were automatically allocated with center as a stratification factor through the study platform to the Retina4IRD-assisted diagnostic arm (AI-assisted group) or specialist-only diagnostic arm without Retina4IRD (control group). Participants assigned to the AI-assisted group underwent basic medical history documentation, fundus photography and OCT examination. The collected data were then input into the Retina4IRD system, which generated top-5 potential genotype predictions for physician review and final confirmation. Participants in the control group underwent the same baseline evaluations. Physicians independently assessed whether patients had IRD and provided their own top-5 genotype predictions without Retina4IRD assistance, based solely on the provided CFP, OCT images and clinical metadata.

All specialists involved this trial were highly experienced, each with over a decade of clinical practice in IRD diagnosis and management. All participants subsequently underwent NGS testing to confirm the causative genotypes, which served as the gold standard for IRD genetic diagnosis.

Outcomes. The primary outcome measure was top-5 gene prediction accuracy, defined as the proportion of cases in which the correct genotype was ranked among the top-5 predictions. Specifically, a ranked list of candidate genotypes was generated for each case by the two arms. A prediction was considered correct if the actual genotype, as confirmed by the subsequent NGS report (the gold standard), fell within the top-5 ranked predictions. Secondary outcomes included the accuracy of top-1 to top-4 gene predictions, defined as the proportion of cases in which the correct genotype prediction was ranked among the top-*N* predictions. A prediction was considered correct if the actual genotype, as confirmed by the subsequent NGS report, fell within the top-*N* ranked predictions. For the detailed formula used to calculate Top-*N* accuracy, see the 'Statistical analysis' section.

We exploratorily assessed the correct revisions proportion (CRP), the incorrect revision proportion (IRP) and the net benefit proportion (NBP) among physicians in the Retina4IRD-assisted group after their review of the Retina4IRD output. CRP is defined as the proportion of cases in which the physician's initial diagnosis was incorrect but was corrected after reviewing the Retina4IRD output. IRP refers to the proportion of cases where the initial diagnosis was correct but became incorrect after review of the Retina4IRD output. NBP is defined as the difference between CRP and IRP. CRP, IRP and NBP are calculated based on the formulas $\mathrm{CRP} = \frac{X_1}{X}$, $\mathrm{IRP} = \frac{Y_1}{Y}$ and $\mathrm{NBP} = \mathrm{CRP} - \mathrm{IRP}$, where X denotes the number of initially incorrect annotations by physicians and X_1 is the number of such correct changed cases assisted by Retina4IRD; Y denotes the number of initially correct annotations by physicians; and Y_1 is the number of such incorrectedly changed cases assisted by Retina4IRD.

Moreover, we also conducted the post hoc analyses for usefulness of Retina4IRD on patient downstream management. We defined a composite score for downstream management strategies, which covers four dimensions: (1) potential eligibility for available therapies or clinical trials, (2) recommended follow-up interval, (3) referral for genetic (reproductive) counseling and (4) referral for syndromic/systemic evaluation. Specifically, we provided the AI with knowledge of clinical and genetic information for IRD genotypes, including therapy/trial availability, inheritance pattern, natural history/progression rate and syndromic/systemic manifestations. In the Retina4IRD-assisted arm, the system generated management recommendations across the four domains; physicians then complete the management report after reviewing the AI output. In the control arm, physicians complete the report without AI assistance. After recommendations were made in both arms, the composite downstream management score was independently assessed by two senior IRD experts under masked conditions using specified criteria (Supplementary Table 25), based on the correct top-1 predicted genotype confirmed by gene sequencing. Each dimension was scored on a 10-point scale, yielding a total score ranging from 0 to 40. A higher score corresponds to a better management plan that better aligns with genotype-specific pathways.

Statistical analysis

Retina4IRD evaluation, for both internal and external validation, was conducted using top-*N* accuracy. The top-*N* accuracy for each class is defined as follows:

$$\mathrm{Accuracy}_c^N = \frac{M_c}{N_c}$$

where $\mathrm{Accuracy}_c^N = \frac{M_c}{N_c}$ is the top-*N* accuracy for class *c*; N_c is the total number of samples in class *c*; and M_c is the number of samples in class *c* for which the true label appears in the model's top-*N* predictions. The overall top-*N* accuracy is defined as follows:

$$\mathrm{Accuracy}_{\mathrm{overall}}^N = \frac{\sum M_c}{\sum N_c}$$

where $\mathrm{Accuracy}_{\mathrm{overall}}^N = \frac{\sum M_c}{\sum N_c}$ is the top-*N* accuracy over all samples; $\sum N_c$ is the total number of samples across all classes; and $\sum M_c$ is the number of total samples for which the true label appears in the model's top-*N* predictions. The model was trained over five independent runs using different random seeds, which controlled the shuffling of the training data. For each run, performance metrics were recorded, and the mean and standard deviation were computed across all runs. The standard error was calculated as the standard deviation divided by $\sqrt{5}$. A 95% CI was then estimated by multiplying the standard error by 1.96. The other measures of model performance included precision and recall. Precision was defined as the proportion of true-positive predictions among all samples predicted as positive: Precision = TP / (TP + FP). Recall was defined as the proportion of actual positive samples correctly identified by the model: Recall = TP / (TP + FN). TP, FP and FN denote true positives, false positives and false negatives, respectively. We also performed exploratory subgroup analyses examining Retina4IRD performance across early-onset, later-onset and end-stage IRD in the China external test dataset ($N = 445$). Early-onset was defined as age at onset no more than 6 years or no more than 12 years; later-onset was defined as age at onset no less than 18 years; and end-stage disease was defined as best-corrected decimal visual acuity no more than 0.1 (ref. 38). Additionally, we conducted a device-stratified analysis (where sample size permitted) in the external test dataset to assess performance differences across major retinal scanner platforms.

In the RCT, a sample size of 300 for statistical power above 95% with a two sided $\alpha = 0.05$ was determined based on the hypothesis that the AI-based CDSS group would achieve 85% accuracy in top-5 gene prediction compared to 60% in the specialist-only group. The calculation was performed using the following formula for comparing two independent proportions:

$$n = \frac{\left(Z_{\alpha/2} + Z_{\beta}\right)^2 \times \left[p_1\left(1 - p_1\right) + p_2\left(1 - p_2\right)\right]}{\left(p_1 - p_2\right)^2}$$

The primary analysis was performed on a full analysis set (FAS) basis ($N = 295$). Hypotheses tested were two-sided, and a P value of less than 0.05 was considered statistically significant. In addition, several post hoc analyses were conducted, including subgroup analyses based on sex (females versus males), comparison of the correct identification among genetically positive cases using the genetic testing report as the gold standard and the difference of composite score for downstream management strategies between the two arms.

All analyses, during both stage 1 and stage 2, were conducted at the patient level. This study is reported in accordance with DECIDE-AI reporting guidelines[39]. Analyses were performed using Python (version 3.8.20).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Data can be shared only for noncommercial academic purposes and will require a formal material transfer agreement. Interested investigators can obtain and certify the data transfer agreement and submit requests to X.S. (xdsun@sjtu.edu.cn). Investigators who consent to the terms of the data transfer agreement, including, but not limited to, the use of these data only for academic purposes, and who agree to protect the confidentiality of the data and limit the possibility of identification of patients will be granted access. Requests will be evaluated on a case-by-case basis within 1 month before receipt of a response. All data shared will be deidentified. For the reproduction of our algorithm code, we have also deposited a minimum dataset at Zenodo (https://zenodo.org/records/20489924) (ref. 40), which is publicly available for scientific research and noncommercial use. Source data are provided with this paper.

Code availability

The code being used in the current study for developing the algorithm is provided at https://github.com/zycl2001/Retina4IRD.

References

37. Lawrence, N. D. A unifying probabilistic perspective for spectral dimensionality reduction: insights and new models. *J. Mach. Learn. Res.* **13**, 1609–1638 (2012).
38. Consultation on development of standards for characterization of vision loss and visual functioning: Geneva, 4–5 September 2003. https://iris.who.int/handle/10665/68601 (World Health Organization, 2003).
39. Vasey, B. et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat. Med.* **28**, 924–933 (2022).
40. Sun, X. An AI-based clinician-decision support system for inherited retinal diseases: a multicenter clinical validation trial. *Zenodo*. https://zenodo.org/records/20489924 (2026).

Acknowledgements

We sincerely thank all of the hospitals for providing IRD data and the participating retinal specialists who generously donated their time and efforts for the reader studies.

Author contributions

X.S. and T.Y.W. initiated the project, organized a collaborative team and provided study supervision. H.J., B.S., X.S. and T.Y.W. conceived of the study and its design. Y.Q., J. Wang, J.C., T.L., G.Z., Y. Jin, H.C., J.X., K.X., Y.R., Y. Jiang, Y.Y., W.L., Q.G., Ya Li, Yang Li, B.L., J. Wu, Q.Z., F.L. and W.Z. collected the data for the model's training and testing. B.S., B.Q., C.Z., S.L., X.C. and D.Z. developed the architectures, training and validation setup. J.H., S.H.B., J.L., A.S, M.S., T.H.R. and A.G. collected the data for the international external validation. X.S., H.J., Yang Li, J. Wu, Q.Z., F.L., W.Z. and Q.G. lead the multicenter RCT. H.J., B.Q. and Y.Q. wrote the paper. X.S., T.Y.W. and B.S. critically revised the paper, and all authors discussed the results and provided feedback regarding the paper. J.H., A.G., Yang Li, J. Wu, Q.Z., F.L., W.Z. and B.L. are the co-leaders of this study and senior authors of this paper. All authors had final responsibility for the decision to submit for publication.

Funding

X.S. discloses support for the research of this work from the National Natural Science Foundation of China (grant 82388101); the National Key R&D Program of China (grant 2022YFC2502800); and the Noncommunicable Chronic Diseases—National Science and Technology Major Project (grants 2023ZD0509202 and 2023ZD0509201). H.J. discloses support for the research of this work from the National Natural Science Foundation of China (grant 82471130); the Shanghai Municipal Talent Work Bureau (grant QNWS2025029); and the Shanghai Jiao Tong University School of Medicine (grant 20250412). B.Q. discloses support for the research of this work from the National Natural Science Foundation of China (grant 62502206). B.S. discloses support for the research of this work from the National Natural Science Foundation of China (grant T2525004) and the Noncommunicable Chronic Diseases—National Science and Technology Major Project (grants 2023ZD0509202 and 2023ZD0509201). D.Z. discloses support for the research of this work from the National Natural Science Foundation of China (grant 62136004) and the National Key R&D Program of China (grant 2023YFF1204803). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Competing interests

T.Y.W. is a consultant for 4MDT, Aldropika Therapeutics, Bayer, Boehringer Ingelheim, Carl Zeiss, Plano, Quaerite Biopharm Research Ltd., Roche and Shanghai Henlius. He is an inventor, holds patents and is a co-founder of the start-up company EyRiS, which has interests in, and develops digital solutions for, eye diseases. X.S. is a consultant for Alcon, Allergan, Bayer Healthcare, Carl Zeiss, Innovent Biologics and Roche. The other authors declare no competing interests.

Additional information

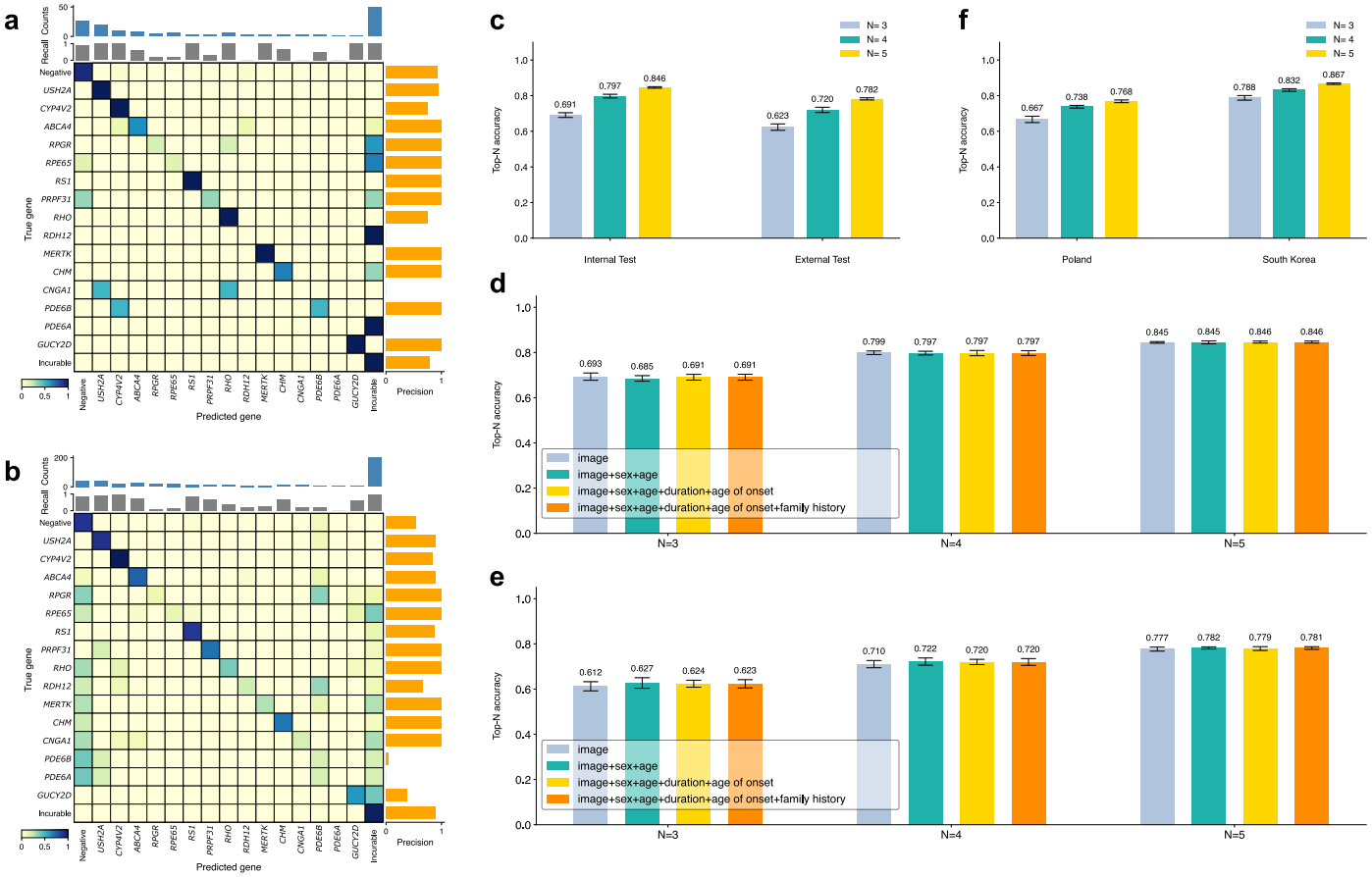
Extended data is available for this paper at https://doi.org/10.1038/s41591-026-04545-w.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-026-04545-w.

Correspondence and requests for materials should be addressed to Bin Sheng, Tien Yin Wong or Xiaodong Sun.

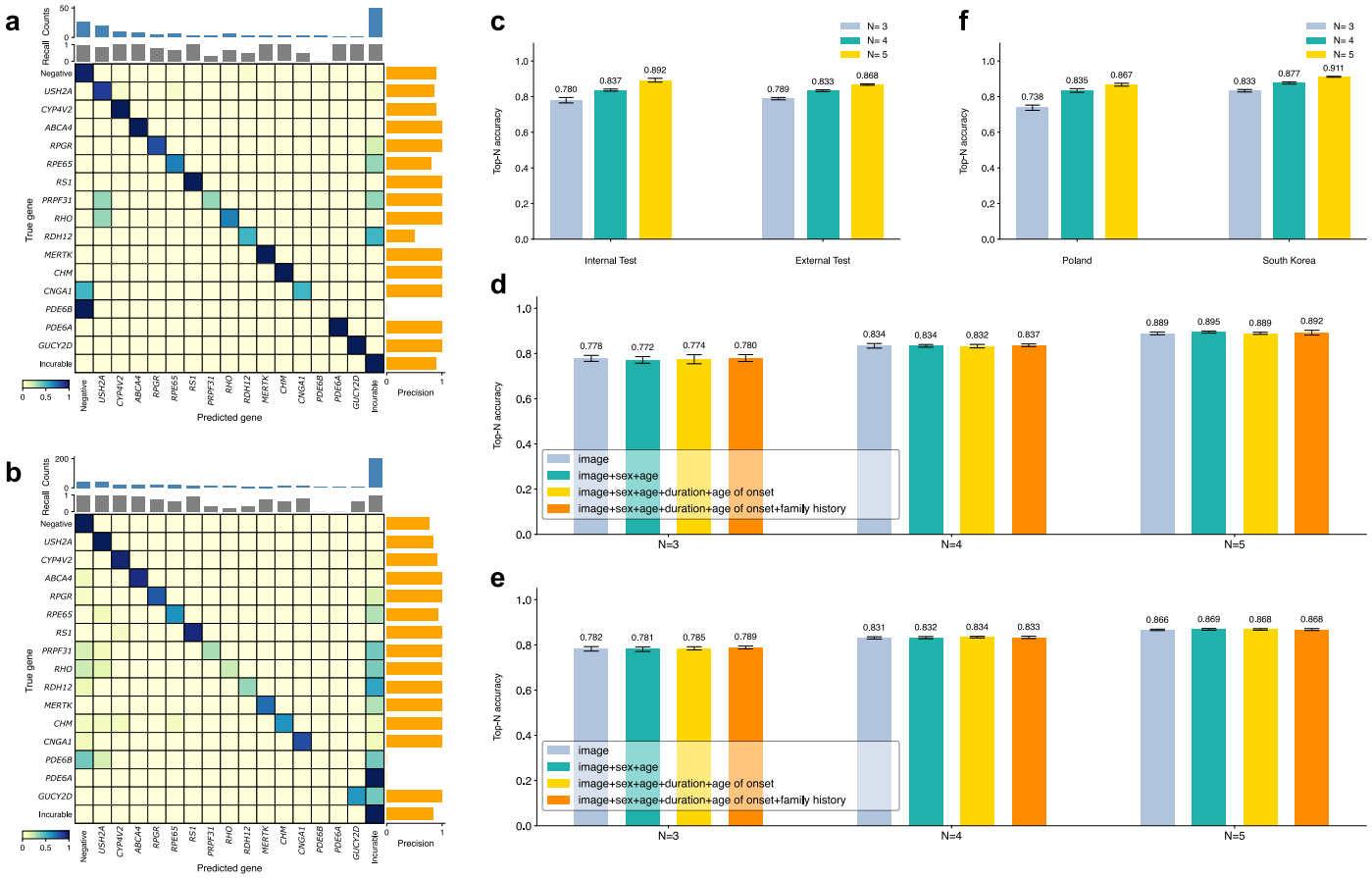
Peer review information *Nature Medicine* thanks Nikolas Pontikos, Gui-shuang Ying and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Mattia Andreoletti, in collaboration with the *Nature Medicine* team.

https://doi.org/10.1038/s41591-026-04545-w



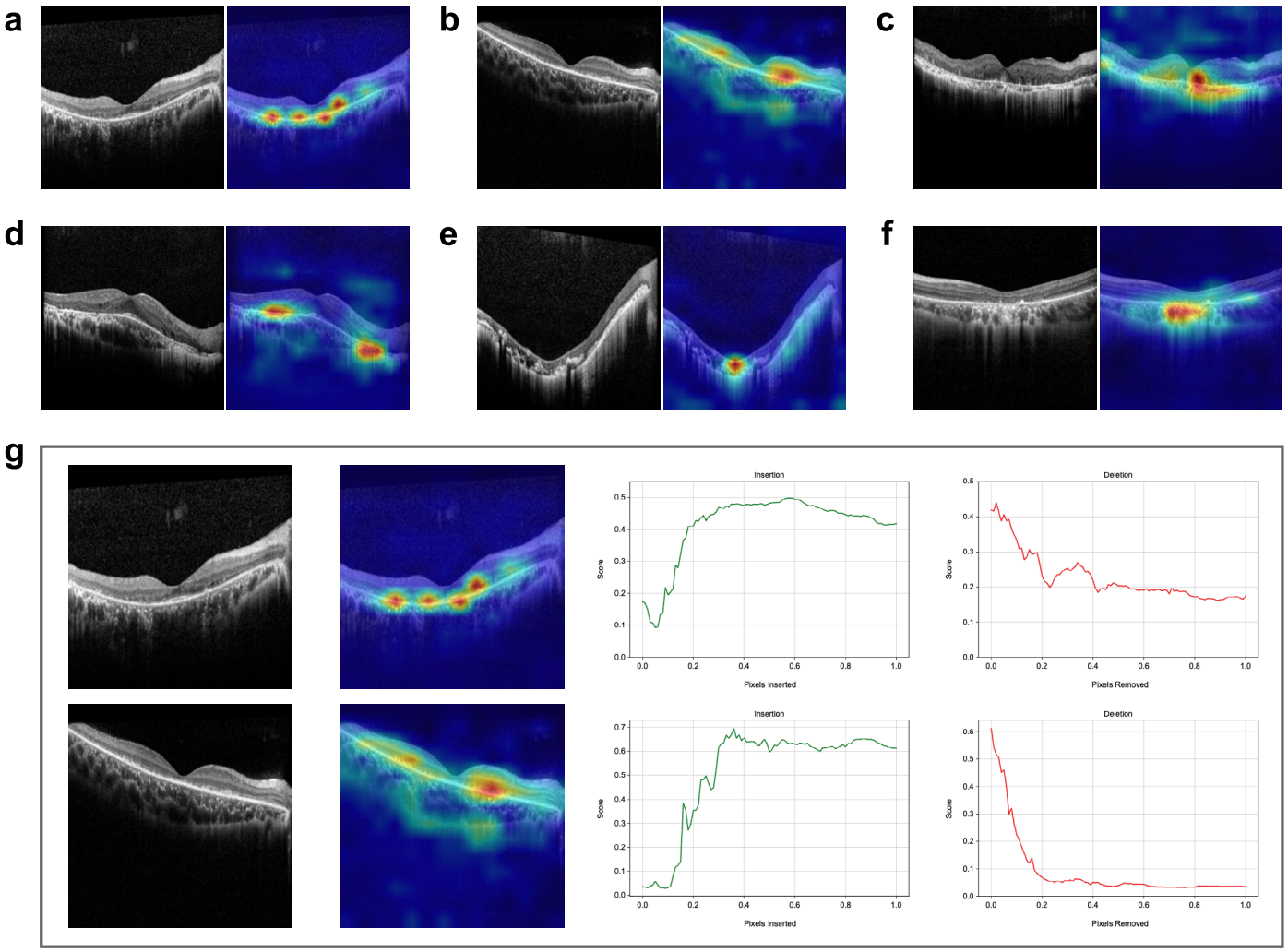
Extended Data Fig. 1 | Performance of Retina4IRD based on CFP and metadata. **a**, Evaluation of the model on the internal test set (n = 148 cases) across 17 classes using CFP images and metadata, including 26 negative cases, 72 potentially curable, and 50 incurable cases. Patient counts, precision, and recall for each class are displayed alongside the confusion matrix. **b**, Evaluation of the model on the external test set (n = 455 cases) across 17 classes using CFP images and metadata, including 42 negative cases, 211 potentially curable, and 202 incurable cases. Patient counts, precision, and recall for each class are displayed alongside the confusion matrix. **c**, Top-N accuracy (N ∈ {3,4,5}) for gene prediction on internal and external test sets based on CFP images and metadata, conducted via 148 and 455 cases, respectively. The bar represents the accuracy and the error bar represents 95% CI. **d**, Top-N accuracy (N ∈ {3,4,5}) for gene prediction using progressive integration of metadata with CFP images on the internal test set (n = 148 cases). The bar represents the accuracy and the error bar represents 95% CI. **e**, Top-N accuracy (N ∈ {3,4,5}) for gene prediction using progressive integration of metadata with CFP images on the external test set (n = 455 cases). The bar represents the accuracy and the error bar represents 95% CI. **f**, Top-N accuracy (N ∈ {3,4,5}) for gene prediction on datasets from Poland (n = 87 cases) and South Korea (n = 164 cases). The bar represents the accuracy and the error bar represents 95% CI.

https://doi.org/10.1038/s41591-026-04545-w



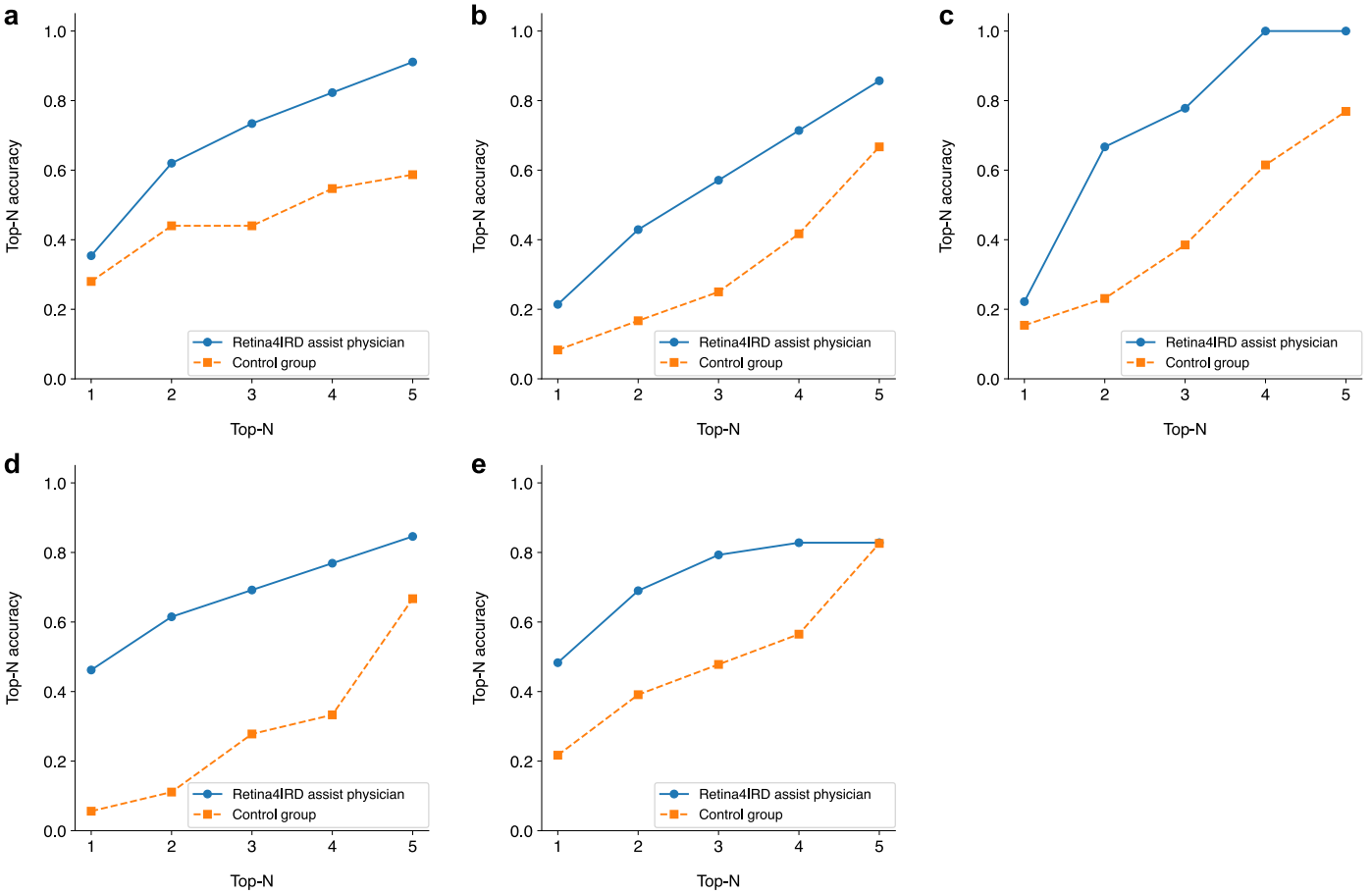
Extended Data Fig. 2 | Performance of Retina4IRD based on OCT and metadata. **a**, Evaluation of the model on the internal test set (n = 148 cases) across 17 classes using OCT images and metadata, including 26 negative cases, 72 potentially curable, and 50 incurable cases. Patient counts, precision, and recall for each class are displayed alongside the confusion matrix. **b**, Evaluation of the model on the external test set (n = 455 cases) across 17 classes using OCT images and metadata, including 42 negative cases, 211 potentially curable, and 202 incurable cases. Patient counts, precision, and recall for each class are displayed alongside the confusion matrix. **c**, Top-N accuracy (N ∈ {3,4,5}) for gene prediction on internal and external test sets based on OCT images and metadata, conducted via 148 and 455 cases, respectively. The bar represents the accuracy and the error bar represents 95% CI. **d**, Top-N accuracy (N ∈ {3,4,5}) for gene prediction using progressive integration of metadata with OCT images on the internal test set (n = 148 cases). The bar represents the accuracy and the error bar represents 95% CI. **e**, Top-N accuracy (N ∈ {3,4,5}) for gene prediction using progressive integration of metadata with OCT images on the external test set (n = 455 cases). The bar represents the accuracy and the error bar represents 95% CI. **f**, Top-N accuracy (N ∈ {3,4,5}) for gene prediction on datasets from Poland (n = 87 cases) and South Korea (n = 164 cases). The bar represents the accuracy and the error bar represents 95% CI.

https://doi.org/10.1038/s41591-026-04545-w



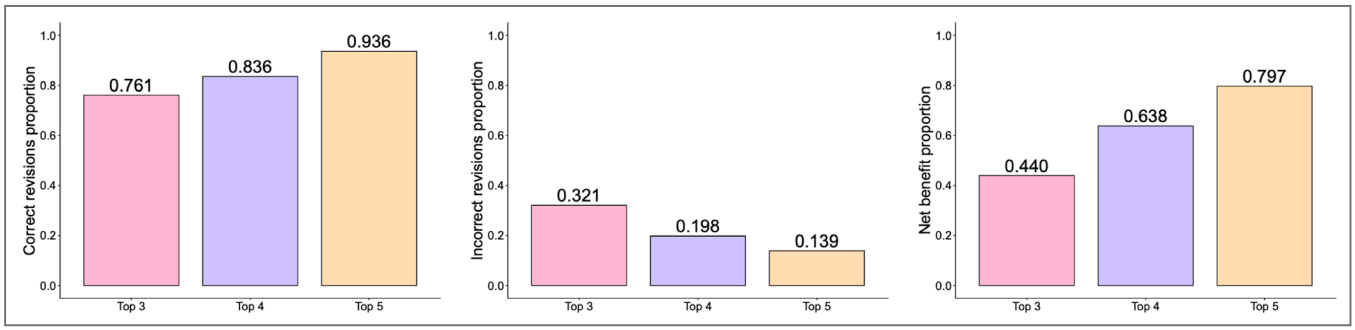
Extended Data Fig. 3 | Visualization and plausibility of Retina4IRD. Panels (**a**)-(**f**) showed the original OCT images and their corresponding attention maps for *USH2A*, *ABCA4*, *CYP4V2*, *CHM*, *RDH12*, and *GUCY2D*, respectively. The attention maps highlight the regions contributing to the model's predictions for the six genotypes. **g**, showed the results of the insertion and deletion experiments performed on the attention maps. The first row shows *USH2A*, and the second row shows *ABCA4*.

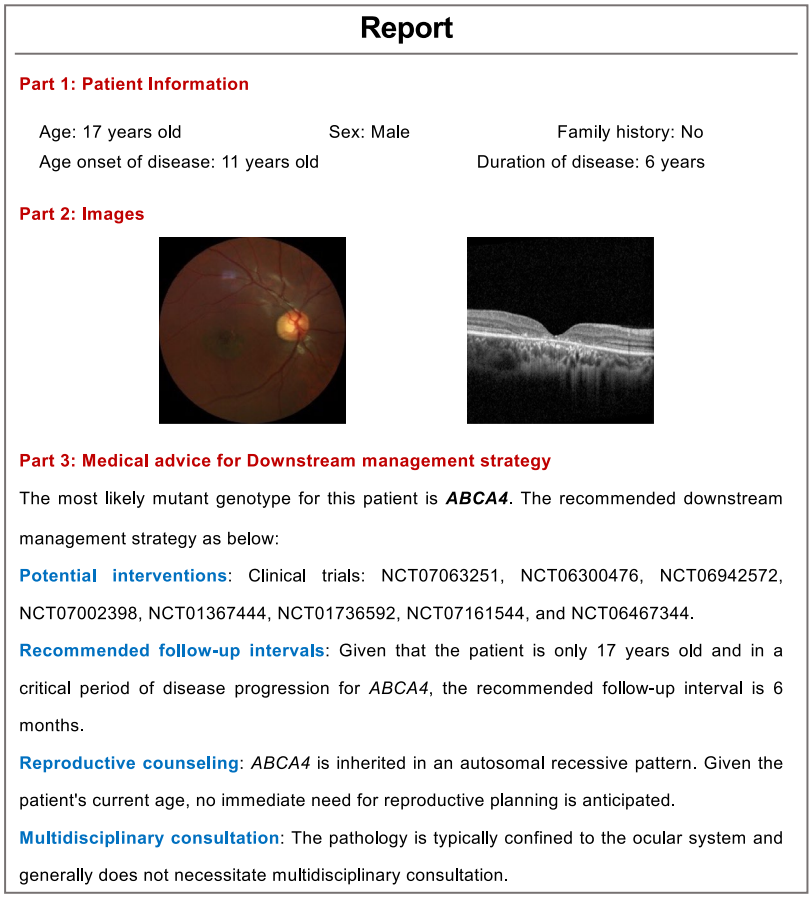
https://doi.org/10.1038/s41591-026-04545-w

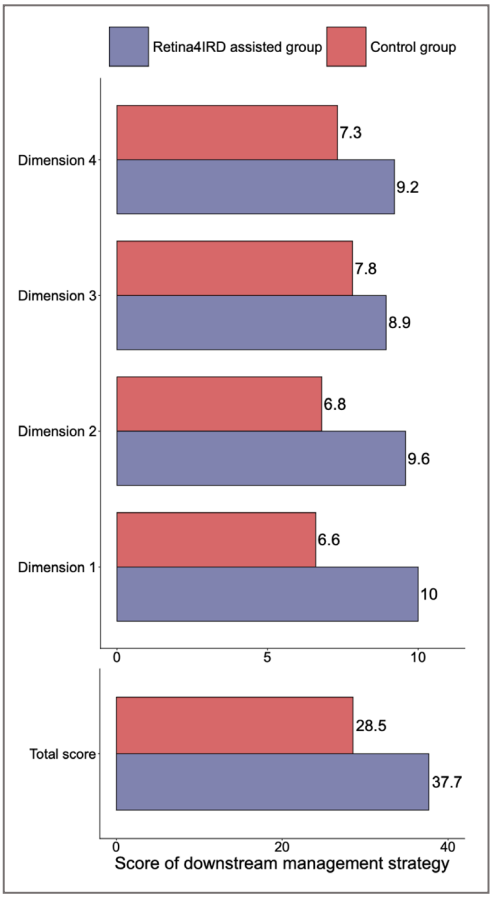


Extended Data Fig. 4 | Top-N Accuracy for each clinical center in FAS. Panels (**a**)-(**e**) present the accuracy for the five main centers in the FAS, labeled as Shanghai General Hospital (n = 154), Eye and ENT Hospital of Fudan University (n = 26), West China Hospital, Sichuan University (n = 22), Beijing Tongren Hospital (n = 31), and Zhongshan Ophthalmic Center, Sun Yat-sen University (n = 52), respectively. The sample sizes for Retina4IRD-assisted group and control group were 79 vs 75 for panel **a**, 14 vs 12 for panel **b**, 9 vs 13 for panel **c**, 13 vs 18 for panel **d**, and 29 vs 23 for panel **e**, respectively. Points represent the accuracy. $N \in \{1,2,3,4,5\}$. FAS, Full Analysis Set.

https://doi.org/10.1038/s41591-026-04545-w







Extended Data Fig. 5 | Retina4IRD enhances physician diagnostic accuracy and downstream management decisions in an RCT. Panels (**a**)-(**c**) present the Correct revisions proportion (CRP), incorrect revision proportion (IRP), and the net benefit proportion (NBP) for physicians after reviewing the Retina4IRD output in the assisted group (n = 148). CRP is defined as the proportion of cases in which the physician's initial diagnosis was incorrect but was corrected after reviewing the Retina4IRD output. IRP refers to the proportion of cases where the initial diagnosis was correct but became incorrect following review of the Retina4IRD output. NBP is defined as the difference between CRP and IRP. The CRP, IRP and NBP are calculated based on the formula: $CRP = \frac{X_1}{X}$, $IRP = \frac{Y_1}{Y}$, and, $NBP = CRP - IRP$, Where the X denotes the number of initially incorrect annotations by doctors and X_1 is the number of such correct changed cases assisted by Retina4IRD; Y denotes the number of initially correct annotations by doctors and Y_1 is the number of such incorrectedly changed cases assisted by Retina4IRD. The bars represented the values of CRP, IRP and NBP for panel **a**, panel **b**, and panel **c** respectively. **d**, Example of a downstream management report in Retina4IRD assisted group. **e**, Mean composite score for downstream management strategies in the physician-only group versus the Retina4IRD-assisted group in the RCT. The composite strategy was developed based on the top-1 predicted genotype and included four dimensions: Dimension 1, potential eligibility for available therapies or clinical trials; Dimension 2, recommended follow-up interval; Dimension 3, referral for genetic (reproductive) counseling; and Dimension 4, referral for syndromic/systemic evaluation. Each dimension was scored on a 10-point scale, yielding a total score ranging from 0 to 40. The composite strategy score was independently assessed by two experienced IRD experts under masked conditions. These recommendations were applied only to cases with a correct top-1 genotype prediction, as confirmed by NGS. A total of 24 management reports from the Retina4IRD-assisted group and 25 management reports from the control group were included in this analysis. The bars represented the averaged scores for total score and Dimension 1 to 4. The *P* values, obtained via the two independent samples t-test, were <0.001, <0.001, <0.001, 0.062, and 0.054, respectively.

nature portfolio

Corresponding author(s): Tien Yin Wong, Bin Sheng, Xiaodong Sun

Last updated by author(s): Jun 21, 2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No special software was used for data collection.
Data analysis	Python version 3.8.20 (Python Software Foundation, Delaware, United States) was used for all computational experiments in this study. The following third-party Python packages were used: PyTorch version 1.10.0 and torchvision version 0.11.1 (Facebook, Menlo Park, California, United States) for deep learning model development; Pandas version 1.3.4 (The Pandas Development Team, United States) for data processing; and Scikit-learn version 1.3.2 (David Cournapeau, California, United States) for calculating accuracy. The code being used in the current study for developing the algorithm is provided at https://github.com/zycl2001/Retina4IRD.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Data can only be shared for non-commercial academic purposes and will require a formal material transfer agreement. Interested investigators can obtain and certify the data transfer agreement and submit requests to X.S. (xdsun@sjtu.edu.cn) Investigators who consent to the terms of the data transfer agreement, including, but not limited to, the use of these data only for academic purposes, and to protect the confidentiality of the data and limit the possibility of identification of patients, will be granted access. Requests will be evaluated on a case-by-case basis within one month before receipt of a response. All data shared will be deidentified. For the reproduction of our algorithm code, we have also deposited a minimum dataset at Zenodo (https://zenodo.org/records/20489924), which is publicly available for scientific research and noncommercial use. Source data are provided with this paper.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	This study strictly adheres to the SAGER guidelines. Consistent with these recommendations, we have applied the term "sex" solely in reference to biological attributes. Participant sex was determined from official identity (ID) documentation, and the data were classified into two categories: Female and Male. Descriptive data analysis was reported by sex is described in Table 1 . Post hoc subgroup analyses according to sex are presented in Figure 5.
Reporting on race, ethnicity, or other socially relevant groupings	For the development and validation of Retina4IRD, multi-ethnic datasets were used. The cohort demographics from different ethnics were shown in Table 1. The performance of Retina4IRD in multi-ethnic datasets were presented in Figure 3, Figure 5, Extended Data Figure 1, and Extended Data Figure 2.
Population characteristics	Detailed population characteristics was given in Table 1.
Recruitment	This is a two-stage study. In Stage 1, we retrospectively collected 5,167 images and 8,901 metadata items from 1,843 genetically confirmed individuals across China, South Korea, and Poland. Multi-ethnic but predominantly East Asian sample may introduce geographic bias. Stage 2 is a randomized controlled validation trial including 300 participants conducted in China to evaluate the effectiveness of Retina4IRD in real-world clinical settings. Self-selection bias may overestimate effectiveness, as volunteered physicians and patients may have stronger interest in AI.
Ethics oversight	This study was conducted in accordance with all applicable ethical regulations. Stage 1, a retrospective model development study, was approved by the Ethics Committee of the National Clinical Research Centre for Eye Diseases (2022KY140, approved December 30, 2022; 2022KY140-2, approved December 29, 2023). Stage 2, a randomized controlled trial, was approved by the same Ethics Committee (2024-451, approved November 6, 2024) and registered with ClinicalTrials.gov (NCT06839170).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	In the RCT, a sample size of 300 for statistical power above 95% with a two side Alpha = 0.05, determining based on the hypothesis that the AI-based CDSS group would achieve 85% accuracy in top-5 gene prediction, compared to 60% in the specialist-only group.
Data exclusions	During screening, 107 cases were excluded (88 declined, 19 met exclusion criteria). In the final analysis, 5 additional participants were excluded for missing NGS test reports. All exclusion criteria were pre-established and applied consistently. Exclusion criteria for the RCT were refusal to undergo NGS genetic testing, fundus photography, or OCT examination; presence of other ocular conditions that may interfere with retinal examination; history of intraocular surgery in either eye within 6 months prior to screening; systemic severe diseases, intellectual disability, or psychiatric disorders; participation in other interventional clinical trials, such as stem cell or gene therapy trials.
Replication	Replication is not relevant. We used independent validation cohorts and experiments to test Retina4IRD, and it achieved similar performance.

Randomization	Randomization was performed in a 1:1 ratio using a computer-generated pseudo-random number sequence.
Blinding	Participated patients and assessors will be masked.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	ClinicalTrials.gov NCT06839170
Study protocol	Protocol was attached in Supplementary.
Data collection	The trial was conducted from February 27, 2025 to July 1, 2025 at Shanghai General Hospital; Eye & ENT Hospital of Fudan University; Beijing Tongren Hospital; Zhongshan Ophthalmic Center, Sun Yat-sen University; West China Hospital, Sichuan University; Henan Provincial People's Hospital; People's Hospital of Ningxia Hui Autonomous Region. Collected data include color fundus photo, OCT image, sex, age, family history, age onset, disease duration and genotype.
Outcomes	The primary outcome measure was top-5 gene prediction accuracy, defined as the proportion of cases in which the correct genotype was ranked among the top 5 predictions. Specifically, a ranked list of candidate genotypes was generated for each case by the two arms. A prediction was considered correct if the actual genotype, as confirmed by the subsequent NGS report (the gold standard), fell within the top-5 ranked predictions. The secondary outcomes included the accuracy of top-1 to top-4 gene predictions, defined as the proportion of cases in which the correct genotype prediction was ranked among the top-N predictions. A prediction was considered correct if the actual genotype, as confirmed by the subsequent NGS report, fell within the top-N ranked predictions.

Plants

Seed stocks	not applicable
Novel plant genotypes	not applicable
Authentication	not applicable