nature medicine

Supplementary information

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

In the format provided by the authors and unedited

SUPPLEMENTARY INFORMATION

Supplementary information - Section A: Supplementary Tables and Figures

Supplementary information - Section B: Study protocols for the randomized controlled trial

Supplementary information - Section A

This supplementary appendix has been provided by the authors to give readers additional information about their work.

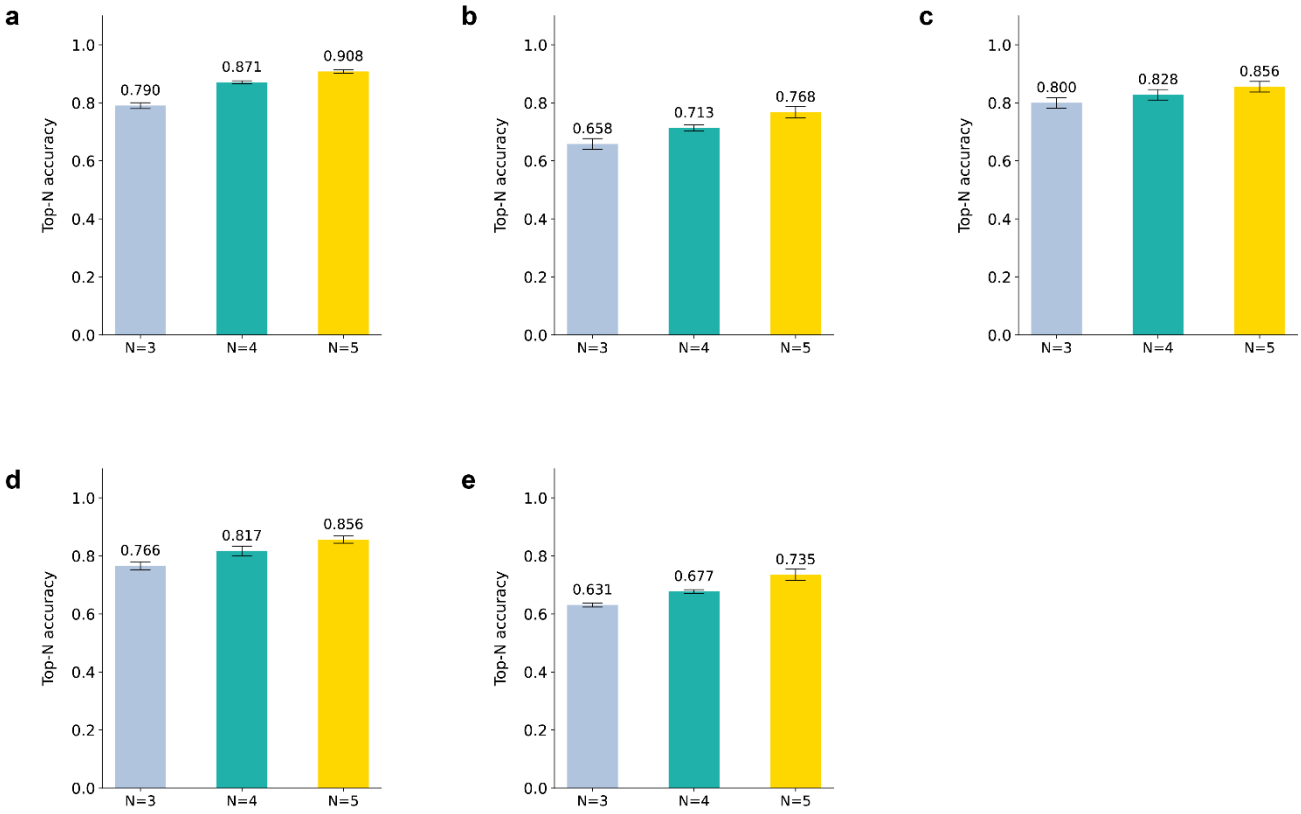

Figure S1. Site-specific top-N accuracy in external test set based on CFP, OCT, and metadata.

Panels (**a**)-(**e**) present the Top-N accuracy in subjects from Shanghai General Hospital (n = 209), the Eye and ENT Hospital of Fudan University (n = 62), West China Hospital, Sichuan University (n = 36), Beijing Tongren Hospital (n = 82), and Zhongshan Ophthalmic Center, Sun Yat-sen University (n = 52), respectively. N ∈ {3,4,5}. The bar represents the accuracy and the error bar represents 95% CI.

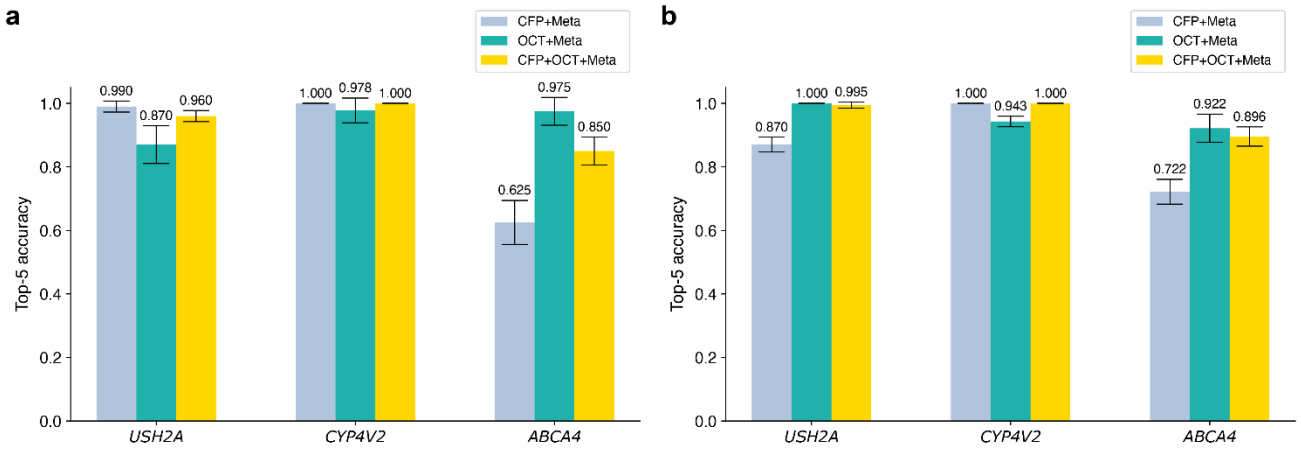

Figure S2. Prediction performance for the genotypes with the top three largest sample sizes.

a, Prediction performance on the internal test set for single genotype. The three genotypes with the highest number of patients were included in this study, namely *USH2A* (n = 20), *CYP4V2* (n = 9), and *ABCA4* (n = 8). The Top-5 accuracy was calculated. The bar represents the accuracy and the error bar represents 95% CI. **b**, Prediction performance on the external test set for single genotype. The three genotypes with the highest number of patients were included in this study, namely *USH2A* (n = 37), *CYP4V2* (n = 21), and *ABCA4* (n = 23). The Top-5 accuracy was calculated. The bar represents the accuracy and the error bar represents 95% CI.

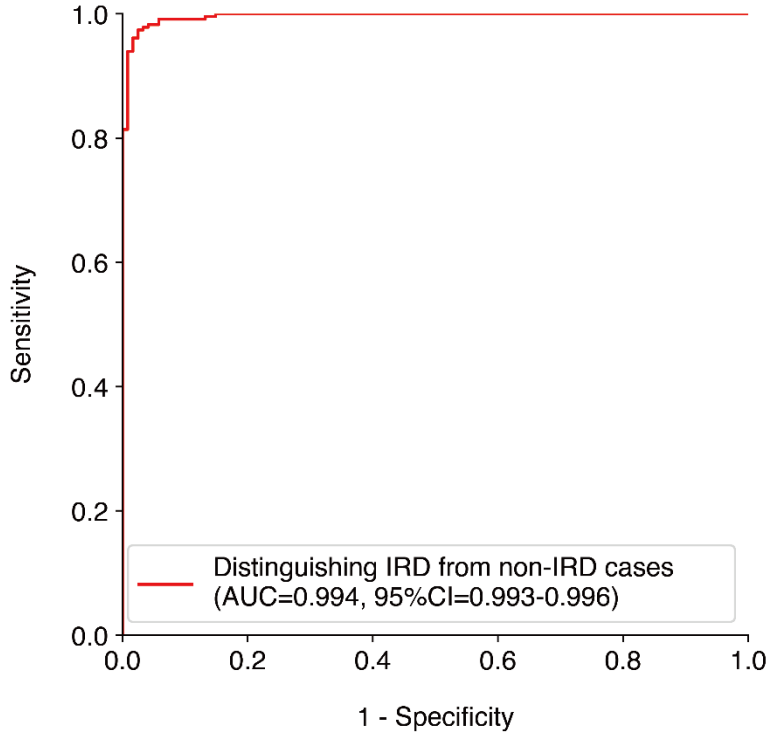

Figure S3. ROC curve of the CFP-only model for distinguishing IRD from non-IRD cases.

This distinguishing model was developed using a cohort of 458 participants, comprising 177 without IRD and 281 with IRD, validated on an independent test set of 194 participants (72 non-IRD and 122 IRD cases), achieving an area under the curve of 0.994 (95% confidence interval: 0.993-0.996). AUC, area under the curve.

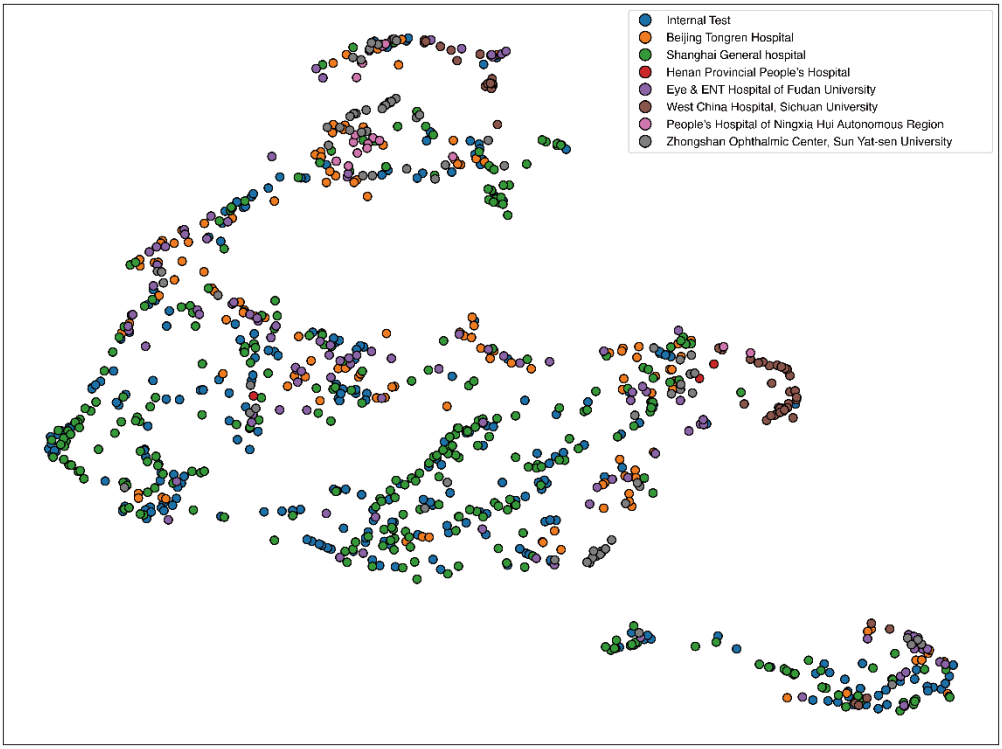

Figure S4. UMAP-based visualization of OCT feature distribution across study centers.

UMAP was employed to visualize the feature embeddings learned by Retina4IRD using a cohort comprising 148 internal samples and 455 external samples collected from seven ophthalmic centers. Samples from different clinical centers exhibited a well-mixed distribution without forming distinct institution-specific clusters. The absence of center-dependent separation suggests that the learned representations are driven primarily by disease-related characteristics rather than institution- or region-specific factors, demonstrating robust generalization across heterogeneous clinical environments.

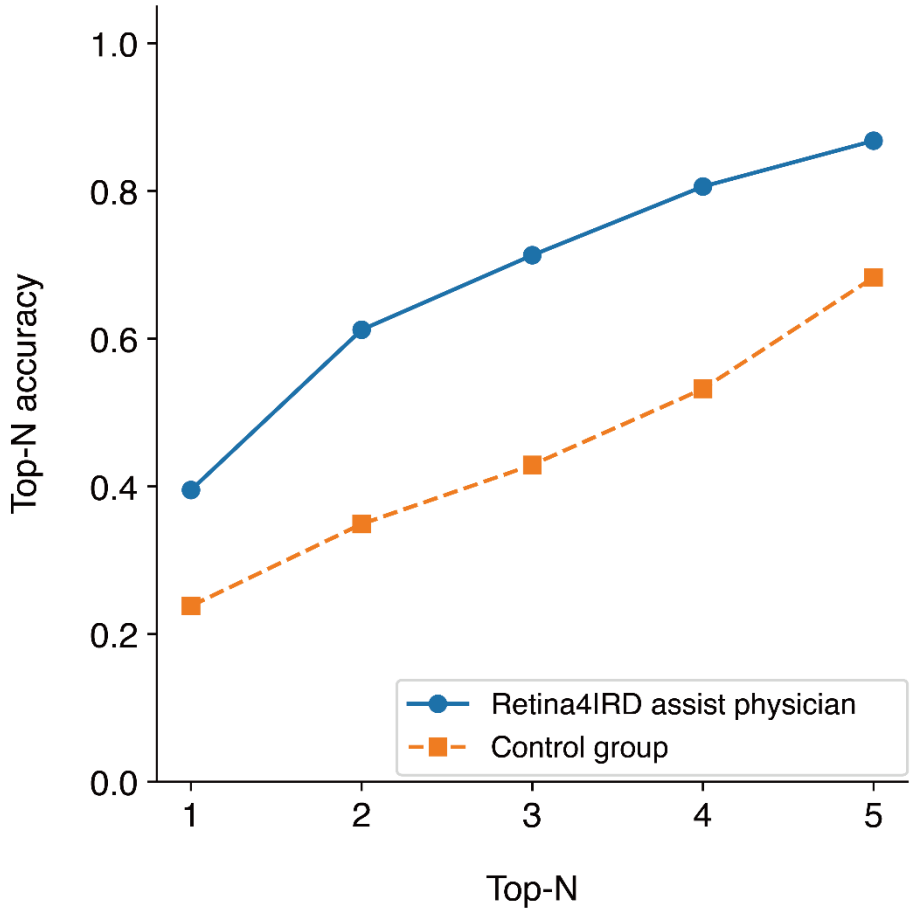

Figure S5. Top-N accuracy among participants with a positive genetic report.

The Top-N accuracy for participants with a positive genetic report was presented, with the sample sizes of 129 for Retina4IRD-assisted group and 126 for control group, respectively. Points represent the accuracy. $N \in \{1,2,3,4,5\}$.

Table S1. The distribution of inherited retinal disease datasets.

Type	Country	City	Hospital	Eyes	Patients
Model training, validation, and internal test dataset	China	Shanghai	Shanghai General Hospital	2197	1137
External test dataset 1	China	Shanghai	Shanghai General Hospital	295	209
	China	Beijing	Beijing Tongren Hospital	136	82
	China	Yinchuan	People's Hospital of Ningxia Hui Autonomous Region	18	11
	China	Shanghai	Eye and ENT Hospital of Fudan University	102	62
	China	Guangzhou	Zhongshan Ophthalmic Center, Sun Yat-sen University	74	52
	China	Chengdu	West China Medical Center, Sichuan University	52	36
	China	Zhengzhou	Henan Provincial People's Hospital	4	3
External test dataset 2	Poland	Olsztyn	University of Warmia and Mazury	172	87
External test dataset 3	South Korea	Seoul	Severance Hospital	326	164
RCT dataset	China	Shanghai	Shanghai General Hospital	154	154
	China	Beijing	Beijing Tongren Hospital	31	31
	China	Yinchuan	People's Hospital of Ningxia Hui Autonomous Region	8	8
	China	Shanghai	Eye and ENT Hospital of Fudan University	26	26
	China	Guangzhou	Zhongshan Ophthalmic Center, Sun Yat-sen University	52	52
	China	Chengdu	West China Medical Center, Sichuan University	22	22
	China	Zhengzhou	Henan Provincial People's Hospital	2	2

Table S2. Baseline characteristics of study population by IRD.

	Stage 1 (N = 1843)		Stage 2 (N = 295) *	
	IRD	**non-IRD**	**IRD**	**non-IRD**
Sample size	1549	294	255	40
Age, years	36 [1, 76]	41 [0, 79]	33 [2, 69]	37 [3, 67]
Sex, n (%)				
Males	958 (61.9)	169 (57.5)	157 (61.6)	24 (60.0)
Females	591 (38.1)	125 (42.5)	98 (38.4)	16 (40.0)
Age onset of disease, years	10 [0, 67]	14 [0, 70]	7 [0, 52]	9 [0, 43]
Duration of disease, years	21 [0, 71]	22 [0, 64]	17 [0, 56]	27 [0, 57]
Family history, n (%)				
No	1247 (82.5)	258 (89.0)	204 (80.0)	33 (82.5)
Yes	265 (17.5)	32 (11.0)	51 (20.0)	7 (17.5)

Baseline characteristics were described as median (range) or n (%), as appropriate. All results are reported as the number of patients.

Family history: “No” means either “recessive”, “sporadic” or “unknown”, “Yes” means “dominant” or “X-linked”.

*In stage 2, a total of 300 participants were randomized, of whom 295 with available NGS reports were included in the final analysis of the RCT.

Table S3. Image devices distribution across training, validation, internal test and external test datasets.

Datasets	CFP								OCT					
	Cases	**Canon** CR-2/CR-2 plus	**KOWA** nonmyd wx	**Topcon** TRC-NW8	**ZEISS** VISUCAM	**ZEISS** Clarus 500	**NIDEK** AFC-330	**CenterVue** DRS	**Cases**	**Heidelberg** Spectralis	**ZEISS** Cirrus HD-OCT	**Topcon** DRI OCT Triton	**SVision** Imaging VG200	**Optopol**
Training	206	0	1	8	87	14	96	0	719	715	0	1	3	0
Validation	63	0	2	0	4	5	52	0	177	177	0	0	0	0
Internal test	148	0	12	0	1	13	122	0	148	146	0	1	1	0
External test	706	141	1	169	10	337	46	2	706	513	105	2	5	81

CFP, color fundus photo; OCT, optical coherence tomography.

Table S4. The patients, recall, precision of the internal test dataset based on CFP, OCT, and metadata.

Class	Patients	Recall	Precision
Negative	26	1.000	0.929
USH2A	20	0.950	0.950
CYP4V2	9	1.000	0.900
ABCA4	8	0.875	1.000
RPGR	5	0.800	1.000
RPE65	6	0.500	0.750
RS1	2	1.000	1.000
PRPF31	3	0.333	1.000
RHO	6	1.000	1.000
RDH12	2	0.500	1.000
MERTK	2	1.000	1.000
CHM	3	1.000	1.000
CNGA1	2	0.000	0.000
PDE6B	2	0.500	1.000
PDE6A	1	1.000	1.000
GUCY2D	1	1.000	1.000
Incurable	50	1.000	0.877

CFP, color fundus photo; OCT, optical coherence tomography.

Table S5. The patients, recall, precision of the external test dataset based on CFP, OCT, and metadata.

Class	Patients	Recall	Precision
Negative	42	1.000	0.646
USH2A	37	1.000	0.902
CYP4V2	21	1.000	0.808
ABCA4	23	0.870	1.000
RPGR	17	0.647	1.000
RPE65	21	0.476	1.000
RS1	15	0.933	0.933
PRPF31	10	0.400	1.000
RHO	13	0.462	1.000
RDH12	9	0.333	1.000
MERTK	7	0.571	1.000
CHM	13	0.692	1.000
CNGA1	10	0.500	1.000
PDE6B	5	0.000	0.000
PDE6A	5	0.000	0.000
GUCY2D	5	0.600	0.750
Incurable	202	1.000	0.898

CFP, color fundus photo; OCT, optical coherence tomography.

Table S6. Top-N accuracy based on CFP, OCT, and metadata.

	Internal Test	External Test
Top-3	0.781 (0.765, 0.797)	0.755 (0.752, 0.757)
Top-4	0.858 (0.850, 0.867)	0.816 (0.811, 0.820)
Top-5	0.904 (0.896, 0.912)	0.856 (0.850, 0.863)

The results were reported as accuracy values with their corresponding 95% confidence intervals.
CFP, color fundus photo; OCT, optical coherence tomography.

Table S7. Ablation study based on CFP, OCT, and metadata in internal test dataset.

	Only image	Image + sex + age	Image + sex + age + duration of disease + age onset of disease	Image + sex + age + duration of disease + age onset of disease + family history
Top-3	0.778 (0.758, 0.799)	0.774 (0.757, 0.792)	0.773 (0.760, 0.786)	0.781 (0.765, 0.797)
Top-4	0.854 (0.842, 0.866)	0.855 (0.846, 0.865)	0.855 (0.848, 0.863)	0.858 (0.850, 0.867)
Top-5	0.899 (0.890, 0.907)	0.904 (0.894, 0.914)	0.901 (0.893, 0.909)	0.904 (0.896, 0.912)

The results were reported as accuracy values with their corresponding 95% confidence intervals.
CFP, color fundus photo; OCT, optical coherence tomography.

Table S8. Ablation study based on CFP, OCT, and metadata in external test dataset.

	Only image	Image + sex + age	Image + sex + age + duration of disease + age onset of disease	Image + sex + age + duration of disease + age onset of disease + family history
Top-3	0.750 (0.747, 0.753)	0.752 (0.746, 0.759)	0.752 (0.747, 0.757)	0.755 (0.752, 0.757)
Top-4	0.814 (0.810, 0.819)	0.811 (0.809, 0.814)	0.818 (0.810, 0.826)	0.816 (0.811, 0.820)
Top-5	0.852 (0.846, 0.859)	0.856 (0.849, 0.863)	0.855 (0.851, 0.860)	0.856 (0.850, 0.863)

The results were reported as accuracy values with their corresponding 95% confidence intervals.
CFP, color fundus photo; OCT, optical coherence tomography.

Table S9. The patients, recall, precision of the internal test dataset based on CFP and metadata.

Class	Patients	Recall	Precision
Negative	26	0.923	0.923
USH2A	20	1.000	0.952
CYP4V2	9	1.000	0.750
ABCA4	8	0.625	1.000
RPGR	5	0.200	1.000
RPE65	6	0.167	1.000
RS1	2	1.000	1.000
PRPF31	3	0.333	1.000
RHO	6	1.000	0.750
RDH12	2	0.000	0.000
MERTK	2	1.000	1.000
CHM	3	0.667	1.000
CNGA1	2	0.000	0.000
PDE6B	2	0.500	1.000
PDE6A	1	0.000	0.000
GUCY2D	1	1.000	1.000
Incurable	50	1.000	0.794

CFP, color fundus photo.

Table S10. The patients, recall, precision of the external test dataset based on CFP and metadata.

Class	Patients	Recall	Precision
Negative	42	0.881	0.544
USH2A	37	0.892	0.892
CYP4V2	21	1.000	0.840
ABCA4	23	0.739	0.895
RPGR	17	0.118	1.000
RPE65	21	0.143	1.000
RS1	15	0.867	0.867
PRPF31	10	0.700	1.000
RHO	13	0.385	1.000
RDH12	9	0.222	0.667
MERTK	7	0.286	1.000
CHM	13	0.692	1.000
CNGA1	10	0.200	1.000
PDE6B	5	0.200	0.037
PDE6A	5	0.000	0.000
GUCY2D	5	0.600	0.375
Incurable	202	0.980	0.888

CFP, color fundus photo.

Table S11. Top-N accuracy based on CFP and metadata.

	Internal Test	External Test
Top-3	0.691 (0.677, 0.704)	0.623 (0.605, 0.641)
Top-4	0.797 (0.787, 0.808)	0.720 (0.705, 0.735)
Top-5	0.846 (0.842, 0.850)	0.782 (0.776, 0.788)

The results were reported as accuracy values with their corresponding 95% confidence intervals.
CFP, color fundus photo.

Table S12. Ablation study based on CFP and metadata in internal test dataset.

	Only image	Image + sex + age	Image + sex + age + duration of disease + age onset of disease	Image + sex + age + duration of disease + age onset of disease + family history
Top-3	0.693 (0.677, 0.709)	0.685 (0.673, 0.697)	0.691 (0.678, 0.703)	0.691 (0.677, 0.704)
Top-4	0.799 (0.791, 0.807)	0.797 (0.789, 0.806)	0.797 (0.786, 0.809)	0.797 (0.787, 0.808)
Top-5	0.845 (0.841, 0.848)	0.845 (0.838, 0.851)	0.846 (0.842, 0.850)	0.846 (0.842, 0.850)

The results were reported as accuracy values with their corresponding 95% confidence intervals.
CFP, color fundus photo.

Table S13. Ablation study based on CFP and metadata in external test dataset.

	Only image	Image + sex + age	Image + sex + age + duration of disease + age onset of disease	Image + sex + age + duration of disease + age onset of disease + family history
Top-3	0.612 (0.592, 0.632)	0.627 (0.604, 0.650)	0.624 (0.608, 0.639)	0.623 (0.605, 0.641)
Top-4	0.710 (0.694, 0.727)	0.722 (0.706, 0.738)	0.720 (0.708, 0.732)	0.720 (0.705, 0.735)
Top-5	0.777 (0.768, 0.786)	0.782 (0.777, 0.787)	0.779 (0.771, 0.788)	0.781 (0.776, 0.788)

The results were reported as accuracy values with their corresponding 95% confidence intervals.
CFP, color fundus photo.

Table S14. The patients, recall, precision of the internal test dataset based on OCT and metadata.

Class	Patients	Recall	Precision
Negative	26	0.962	0.893
USH2A	20	0.850	0.850
CYP4V2	9	1.000	0.900
ABCA4	8	1.000	1.000
RPGR	5	0.800	1.000
RPE65	6	0.667	0.800
RS1	2	1.000	1.000
PRPF31	3	0.333	1.000
RHO	6	0.667	1.000
RDH12	2	0.500	0.500
MERTK	2	1.000	1.000
CHM	3	1.000	1.000
CNGA1	2	0.500	1.000
PDE6B	2	0.000	0.000
PDE6A	1	1.000	1.000
GUCY2D	1	1.000	1.000
Incurable	50	1.000	0.893

OCT, optical coherence tomography.

Table S15. The patients, recall, precision of the external test dataset based on OCT and metadata.

Class	Patients	Recall	Precision
Negative	42	1.000	0.764
USH2A	37	1.000	0.841
CYP4V2	21	0.952	0.909
ABCA4	23	0.913	1.000
RPGR	17	0.765	1.000
RPE65	21	0.619	0.929
RS1	15	0.933	1.000
PRPF31	10	0.300	1.000
RHO	13	0.231	1.000
RDH12	9	0.333	1.000
MERTK	7	0.714	1.000
CHM	13	0.615	1.000
CNGA1	10	0.800	1.000
PDE6B	5	0.000	0.000
PDE6A	5	0.000	0.000
GUCY2D	5	0.600	1.000
Incurable	202	1.000	0.845

OCT, optical coherence tomography.

Table S16. Top-N accuracy based on OCT and metadata.

	Internal Test	External Test
Top-3	0.780 (0.765, 0.795)	0.789 (0.782, 0.795)
Top-4	0.837 (0.831, 0.842)	0.833 (0.828, 0.838)
Top-5	0.892 (0.881, 0.903)	0.868 (0.863, 0.872)

The results were reported as accuracy values with their corresponding 95% confidence intervals.
OCT, optical coherence tomography.

Table S17. Ablation study based on OCT and metadata in internal test dataset.

	Only image	Image + sex + age	Image + sex + age + duration of disease + age onset of disease	Image + sex + age + duration of disease + age onset of disease + family history
Top-3	0.778 (0.765, 0.792)	0.772 (0.757, 0.786)	0.774 (0.754, 0.794)	0.780 (0.765, 0.795)
Top-4	0.834 (0.824, 0.844)	0.834 (0.828, 0.840)	0.832 (0.825, 0.840)	0.837 (0.831, 0.842)
Top-5	0.889 (0.883, 0.895)	0.895 (0.890, 0.899)	0.889 (0.885, 0.894)	0.892 (0.881, 0.903)

The results were reported as accuracy values with their corresponding 95% confidence intervals.
OCT, optical coherence tomography.

Table S18. Ablation study based on OCT and metadata in external test dataset.

	Only image	Image + sex + age	Image + sex + age + duration of disease + age onset of disease	Image + sex + age + duration of disease + age onset of disease + family history
Top-3	0.782 (0.772, 0.793)	0.781 (0.771, 0.791)	0.785 (0.777, 0.792)	0.789 (0.782, 0.795)
Top-4	0.831 (0.826, 0.836)	0.832 (0.828, 0.837)	0.834 (0.831, 0.838)	0.833 (0.828, 0.838)
Top-5	0.866 (0.863, 0.870)	0.869 (0.865, 0.873)	0.868 (0.865, 0.872)	0.868 (0.863, 0.872)

The results were reported as accuracy values with their corresponding 95% confidence intervals.
OCT, optical coherence tomography.

Table S19. Prediction performance for the top-5 genotypes in internal test dataset.

Class	CFP + metadata	OCT + metadata	CFP + OCT + metadata
Negative	0.954 (0.929, 0.979)	0.992 (0.979, 1.000)	0.992 (0.979, 1.000)
USH2A	0.990 (0.973, 1.000)	0.870 (0.811, 0.929)	0.960 (0.943, 0.978)
CYP4V2	1.000 (1.000, 1.000)	0.978 (0.939, 1.000)	1.000 (1.000, 1.000)
ABCA4	0.625 (0.556, 0.694)	0.975 (0.931, 1.000)	0.850 (0.806, 0.894)
RPGR	0.480 (0.270, 0.690)	0.680 (0.594, 0.766)	0.680 (0.594, 0.766)
RPE65	0.100 (0.028, 0.172)	0.667 (0.667, 0.667)	0.500 (0.500, 0.500)
RS1	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
PRPF31	0.133 (0.000, 0.277)	0.200 (0.057, 0.343)	0.200 (0.057, 0.343)
RHO	1.000 (1.000, 1.000)	0.700 (0.557, 0.843)	0.967 (0.908, 1.000)
RDH12	0.000 (0.000, 0.000)	0.400 (0.225, 0.575)	0.300 (0.085, 0.515)
MERTK	0.700 (0.485, 0.915)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
CHM	0.600 (0.483, 0.717)	0.800 (0.566, 1.000)	0.867 (0.724, 1.000)
CNGA1	0.000 (0.000, 0.000)	0.400 (0.072, 0.728)	0.200 (0.000, 0.551)
PDE6B	0.500 (0.500, 0.500)	0.000 (0.000, 0.000)	0.500 (0.500, 0.500)
PDE6A	0.000 (0.000, 0.000)	1.000 (1.000, 1.000)	0.600 (0.171, 1.000)
GUCY2D	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
Incurable	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)

The results were reported as accuracy values with their corresponding 95% confidence intervals.
CFP, color fundus photo; OCT, optical coherence tomography.

Table S20. Prediction performance for the top-5 genotypes in external test dataset.

Class	CFP + metadata	OCT + metadata	CFP + OCT + metadata
Negative	0.952 (0.918, 0.987)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
USH2A	0.870 (0.847, 0.894)	1.000 (1.000, 1.000)	0.995 (0.985, 1.000)
CYP4V2	1.000 (1.000, 1.000)	0.943 (0.926, 0.960)	1.000 (1.000, 1.000)
ABCA4	0.722 (0.683, 0.761)	0.922 (0.877, 0.966)	0.896 (0.865, 0.926)
RPGR	0.247 (0.142, 0.352)	0.682 (0.641, 0.724)	0.635 (0.597, 0.674)
RPE65	0.171 (0.115, 0.228)	0.657 (0.596, 0.719)	0.524 (0.487, 0.561)
RS1	0.853 (0.785, 0.922)	0.920 (0.897, 0.943)	0.947 (0.923, 0.970)
PRPF31	0.340 (0.160, 0.521)	0.300 (0.176, 0.424)	0.320 (0.254, 0.386)
RHO	0.385 (0.311, 0.459)	0.385 (0.272, 0.497)	0.477 (0.427, 0.527)
RDH12	0.267 (0.219, 0.314)	0.333 (0.333, 0.333)	0.311 (0.238, 0.384)
MERTK	0.286 (0.207, 0.365)	0.486 (0.336, 0.636)	0.400 (0.306, 0.494)
CHM	0.677 (0.650, 0.704)	0.646 (0.555, 0.738)	0.677 (0.578, 0.776)
CNGA1	0.180 (0.145, 0.215)	0.700 (0.589, 0.811)	0.520 (0.333, 0.707)
PDE6B	0.040 (0.000, 0.110)	0.160 (0.029, 0.291)	0.000 (0.000, 0.000)
PDE6A	0.000 (0.000, 0.000)	0.080 (0.000, 0.220)	0.000 (0.000, 0.000)
GUCY2D	0.400 (0.289, 0.511)	0.520 (0.380, 0.660)	0.440 (0.309, 0.571)
Incurable	0.988 (0.976, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)

The results were reported as accuracy values with their corresponding 95% confidence intervals.
CFP, color fundus photo; OCT, optical coherence tomography.

Table S21. External validation using data from Poland and South Korea.

	Poland	South Korea
CFP + metadata		
Top-3	0.667 (0.649, 0.685)	0.788 (0.775, 0.801)
Top-4	0.738 (0.730, 0.746)	0.832 (0.825, 0.839)
Top-5	0.768 (0.760, 0.775)	0.867 (0.863, 0.871)
OCT + metadata		
Top-3	0.738 (0.723, 0.753)	0.833 (0.826, 0.840)
Top-4	0.835 (0.824, 0.845)	0.877 (0.871, 0.883)
Top-5	0.867 (0.859, 0.875)	0.911 (0.908, 0.914)
CFP + OCT + metadata		
Top-3	0.710 (0.693, 0.728)	0.828 (0.818, 0.838)
Top-4	0.761 (0.738, 0.784)	0.870 (0.857, 0.882)
Top-5	0.825 (0.818, 0.833)	0.901 (0.895, 0.908)

The results were reported as accuracy values with their corresponding 95% confidence intervals.
CFP, color fundus photo; OCT, optical coherence tomography.

Table S22. Retina4IRD performance for the early-onset, late-onset, and end-stage IRD in external test dataset.

Top N accuracy*	Early-onset IRD[#] (n=169)	Early-onset IRD[†] (n=281)	Late-onset IRD (n=108)	End-stage IRD (n=106)
Top-3	0.656 (0.647, 0.665)	0.705 (0.699, 0.712)	0.830 (0.821, 0.838)	0.638 (0.623, 0.653)
Top-4	0.725 (0.718, 0.733)	0.767 (0.761, 0.772)	0.865 (0.857, 0.873)	0.691 (0.675, 0.708)
Top-5	0.795 (0.781, 0.809)	0.821 (0.813, 0.830)	0.880 (0.868, 0.891)	0.736 (0.716, 0.756)

*The performance of Retina4IRD based on CFP, OCT, and metadata in external test dataset were presented as top-n predict accuracy and 95%CI. [#] age of onset ≤6 years; [†]age of onset ≤12 years.

Table S23. The performance of Retina4IRD on different device subgroups in external test dataset.

Top N accuracy	Heidelberg Spectralis (n=513)	ZEISS Cirrus HD-OCT (n=105)	Other platforms (n=88)
Top-3	0.783 (0.777, 0.788)	0.749 (0.739, 0.758)	0.693 (0.668, 0.718)
Top-4	0.843 (0.836, 0.849)	0.781 (0.773, 0.789)	0.748 (0.715, 0.780)
Top-5	0.880 (0.873, 0.887)	0.819 (0.803, 0.835)	0.816 (0.808, 0.824)

Note: The performance of Retina4IRD based on retina image and metadata in external dataset were presented as top-n predict accuracy and 95%CI. Other devices included Topcon DRI OCT Triton, VG200, SVision Imaging, and SOCT REVO FC, Optopol.

Table S24. Summary of the clinical trials for potentially curable IRD.

Therapeutic Strategy	NCT Number	Interventions	Target Gene	Phases	Study Status	Start Date	Completion Date
	NCT06302608	NGGT001	*CYP4V2*	1	ACTIVE_NOT_RECRUITING	2023-02-01	2028-05-01
	NCT06706427	NGGT001	*CYP4V2*	1/2	ACTIVE_NOT_RECRUITING	2024-03-01	2029-09-01
	NCT04722107	rAAV2/8-hCYP4V2	*CYP4V2*	1	UNKNOWN	2021-04-01	2024-04-01
	NCT06699108	VGR-R01	*CYP4V2*	3	NOT_YET_RECRUITING	2024-12-01	2027-06-01
	NCT05399069	VGR-R01	*CYP4V2*	1	UNKNOWN	2022-09-01	2024-05-01
	NCT05714904	ZVS101e	*CYP4V2*	1	UNKNOWN	2022-09-01	2025-03-01
	NCT05832684	ZVS101e	*CYP4V2*	1/2	ACTIVE_NOT_RECRUITING	2023-02-01	2028-12-01
	NCT06743646	ZVS101e	*CYP4V2*	3	RECRUITING	2024-12-01	2030-06-01
	NCT07307469	ZVS101e	*CYP4V2*	1/2	ENROLLING_BY_INVITATION	2026-03-01	2027-07-01
	NCT05694598	VGR-R01	*CYP4V2*	1	UNKNOWN	2023-03-01	2025-09-01
	NCT05858983	FT-001	*RPE65*	1/2	RECRUITING	2022-11-01	2029-11-01
	NCT06196827	LX101	*RPE65*	1/2	ACTIVE_NOT_RECRUITING	2022-07-01	2027-12-01
	NCT07054632	LX101	*RPE65*	3	ACTIVE_NOT_RECRUITING	2023-09-01	2029-08-01
Gene Augmentation	NCT00643747	tgAAG76	*RPE65*	1/2	COMPLETED	2007-01-01	2014-12-01
	NCT02781480	AAV2/5-OPTIRPE65	*RPE65*	1/2	COMPLETED	2016-04-01	2018-12-01
	NCT06088992	HG004	*RPE65*	1	ACTIVE_NOT_RECRUITING	2023-01-01	2028-10-01
	NCT05906953	HG004	*RPE65*	1/2	RECRUITING	2023-10-01	2025-12-01
	NCT06024057	LX101	*RPE65*		NOT_YET_RECRUITING	2023-09-01	2038-09-01
	NCT01496040	rAAV2/4.hRPE65	*RPE65*	1/2	COMPLETED	2011-09-01	2014-08-01
	NCT00749957	rAAV2-CB-hRPE65	*RPE65*	1/2	COMPLETED	2009-06-01	2017-09-01
	NCT00481546	rAAV2-CBSB-hRPE65	*RPE65*	1	ACTIVE_NOT_RECRUITING	2007-07-01	2026-06-01
	NCT00821340	rAAV2-hRPE65	*RPE65*	1	COMPLETED	2009-02-01	2017-01-01
	NCT00516477	Voretigene neparvovec-rzyl	*RPE65*	1	COMPLETED	2007-09-01	2018-03-01
	NCT01208389	Voretigene neparvovec-rzyl	*RPE65*	1/2	ACTIVE_NOT_RECRUITING	2010-11-01	2027-10-01
	NCT00999609	Voretigene neparvovec-rzyl	*RPE65*	3	ACTIVE_NOT_RECRUITING	2012-10-01	2030-01-01
	NCT03920007	ATSN-101	*GUCY2D*	1/2	ACTIVE_NOT_RECRUITING	2019-09-01	2027-05-01
	NCT06749639	PUMCH-E101	*RDH12*	1	RECRUITING	2024-09-01	2030-01-01

NCT06291935	VG901	*CNGA1*	1	RECRUITING	2023-09-01	2026-04-01
NCT01482195	rAAV2-VMD2-hMERTK	*MERTK*	1	COMPLETED	2011-08-01	2019-08-01
NCT04611503	rAAV.hPDE6A	*PDE6A*	1/2	ACTIVE_NOT_RECRUITING	2019-09-01	2027-07-01
NCT03328130	AAV2/5-hPDE6B	*PDE6B*	1/2	RECRUITING	2017-11-01	2029-12-01
NCT07063251	HG005	*ABCA4*	1	RECRUITING	2025-08-01	2028-02-01
NCT06300476	JWK006	*ABCA4*	1/2	ACTIVE_NOT_RECRUITING	2023-11-01	2029-12-01
NCT06942572	SB-007	*ABCA4*	1/2	RECRUITING	2025-02-01	2028-12-01
NCT07002398	VG801	*ABCA4*	1/2	RECRUITING	2024-12-01	2026-05-01
NCT01367444	SAR422459	*ABCA4*	1/2	TERMINATED	2011-06-08	2019-08-16
NCT01736592	SAR422459	*ABCA4*	1/2	ACTIVE_NOT_RECRUITING	2012-12-14	2033-08-29
NCT07161544	AAVB-039	*ABCA4*	1/2	RECRUITING	2025-09-29	2032-07-01
NCT04483440	4D-110	*CHM*	1	ACTIVE_NOT_RECRUITING	2020-06-01	2027-08-01
NCT02407678	AAV2.REP1	*CHM*	2	COMPLETED	2016-08-01	2021-07-01
NCT02341807	AAV2-hCHM	*CHM*	1/2	COMPLETED	2015-01-01	2022-10-01
NCT03507686	BIIB111	*CHM*	2	COMPLETED	2017-11-01	2022-06-01
NCT03496012	BIIB111	*CHM*	3	COMPLETED	2017-12-01	2020-12-01
NCT01461213	rAAV2.REP1	*CHM*	1/2	COMPLETED	2011-10-01	2017-10-01
NCT02077361	rAAV2.REP1	*CHM*	1/2	COMPLETED	2015-04-01	2022-05-01
NCT02553135	rAAV2.REP1	*CHM*	2	COMPLETED	2015-09-01	2018-02-01
NCT02671539	rAAV2.REP1	*CHM*	2	COMPLETED	2016-01-01	2018-02-01
NCT03584165	BIIB111/BIIB112	*CHM*	3	ENROLLING_BY_INVITATION	2018-06-01	2026-06-01
NCT04517149	4D-125	*RPGR*	1/2	ACTIVE_NOT_RECRUITING	2020-06-01	2029-05-01
NCT06646289	AAV5-hRKp.RPGR	*RPGR*	2	RECRUITING	2024-10-01	2030-10-01
NCT04671433	AAV5-hRKp.RPGR	*RPGR*	3	COMPLETED	2020-12-01	2024-09-01
NCT04794101	AAV5-hRKp.RPGR	*RPGR*	3	ACTIVE_NOT_RECRUITING	2020-12-01	2029-12-01
NCT05926583	AAV5-hRKp.RPGR	*RPGR*	3	RECRUITING	2023-09-12	2030-11-12
NCT03252847	AAV5-hRKp.RPGR	*RPGR*	1/2	COMPLETED	2017-07-01	2021-11-01
NCT07174726	AGTC-501	*RPGR*	2	RECRUITING	2025-09-01	2030-12-01
NCT06275620	AGTC-501	*RPGR*	2	ENROLLING_BY_INVITATION	2023-11-01	2029-12-01
NCT03116113	BIIB112	*RPGR*	1/2	COMPLETED	2017-03-01	2020-11-01

	NCT05874310	FT-002	*RPGR*	1	RECRUITING	2023-02-01	2027-11-01
	NCT06492850	FT-002	*RPGR*	1/2	RECRUITING	2024-04-01	2026-02-01
	NCT04850118	rAAV2tYF-GRK1-hRPGRco	*RPGR*	2/3	ACTIVE_NOT_RECRUITING	2024-03-01	2029-10-01
	NCT03316560	rAAV2tYF-GRK1-RPGR	*RPGR*	1/2	ACTIVE_NOT_RECRUITING	2018-04-01	2025-03-01
	NCT06333249	rAAV2tYF-GRK1-RPGR	*RPGR*	2	ACTIVE_NOT_RECRUITING	2021-04-01	2027-02-01
	NCT05878860	ATSN-201	*RS1*	1/2	RECRUITING	2023-08-01	2029-10-01
	NCT06289452	IVB102	*RS1*	1	ACTIVE_NOT_RECRUITING	2024-03-01	2029-12-01
	NCT06345898	JWK002	*RS1*	1	RECRUITING	2023-11-01	2033-11-01
	NCT02416622	rAAV2tYF-CB-hRS1	*RS1*	1/2	COMPLETED	2015-05-01	2023-05-01
	NCT02317887	RS1	*RS1*	1/2	COMPLETED	2015-02-01	2024-10-01
	NCT06066008	ZM-01	*RS1*	1	RECRUITING	2022-09-01	2027-10-01
Gene Editing	NCT05805007	ZVS203e	*RHO*	1	RECRUITING	2023-09-01	2026-04-01
	NCT06952842	ZVS203e	*RHO*	1/2	NOT_YET_RECRUITING	2025-05-18	2045-06-18
Gene Expression Modulation	NCT06388200	OCU400	*RHO*	3	RECRUITING	2024-06-01	2026-12-01
	NCT04123626	QR-1123	*RHO*	1/2	ACTIVE_NOT_RECRUITING	2019-10-01	2022-06-01
	NCT05203939	OCU400	*RHO*	1/2	ACTIVE_NOT_RECRUITING	2022-01-01	2027-03-01
	NCT03780257	QR-421a	*USH2A*	1/2	COMPLETED	2019-03-01	2021-10-01
	NCT05085964	QR-421a	*USH2A*	2	TERMINATED	2021-09-01	2022-10-01
	NCT06627179	QR-421a	*USH2A*	2	RECRUITING	2024-12-01	2027-12-01
	NCT05158296	QR-421a	*USH2A*	2/3	TERMINATED	2021-12-01	2022-10-01
	NCT05176717	QR-421a	*USH2A*	2/3	TERMINATED	2021-12-01	2022-08-01
	NCT06852963	VP-001	*PRPF31*	1/2	RECRUITING	2025-06-01	2028-03-03
	NCT05902962	VP-001	*PRPF31*	1	ACTIVE_NOT_RECRUITING	2023-04-20	2025-08-30
	NCT06455826	VP-001	*PRPF31*	1	RECRUITING	2024-06-13	2025-11-01
	NCT06467344	ACDN-01	*ABCA4*	1/2	RECRUITING	2024-06-01	2030-12-01

Table S25. Scoring criteria and detailed description of assessments for downstream management recommendations.

Domain	Scoring Criteria
1. Potential eligibility for available therapies or clinical trials (Option: 0, 5, 10 points)	
Potential eligibility for available therapies or clinical trials (option: 0, 5, 10)	10 = Correct and comprehensive 5 = Partially correct, with omissions but no incorrect entries 0 = Contains incorrect entries
2. Recommended follow-up interval (Range: 0-10 points)	
Accuracy of follow-up interval (range: 0~10)	10 = Completely consistent with expert judgment 0 = Completely inconsistent with expert judgment or omission
3. Referral for genetic (reproductive) counseling (Range: 0-10 points; If the inheritance pattern is incorrect, the point of this dimension is 0.)	
Inheritance pattern (option: 0 or 3)	3 = Correct; 0 = Incorrect or missing.
Partner-related counseling (range: 0~3)	3 = Completely consistent with expert judgment 0 = Completely inconsistent with expert judgment or missing
Offspring risk counseling (range: 0~3)	3 = Completely consistent with expert judgment 0 = Completely inconsistent with expert judgment or missing
Professionalism of the description (option: 0 or 1)	1 = Clear; 0 = Ambiguous or misleading or incorrect
4. Referral for syndromic/systemic evaluation (0-10 points)	
Judgment on necessity (range: 0~8)	8 = Completely consistent with expert judgment 0 = Completely inconsistent with expert judgment or missing
Specification of specialties/rationale (range: 0~2)	2 = Explicit; 0 = Ambiguous, misleading, or incorrect
Total Score (range: 0~40)	Sum of four domain scores, higher scores indicate better overall report quality.

Note: This management recommendations were applied only to cases with a correct top-1 genotype prediction, as confirmed by NGS.

Explanation of the specified scoring criteria

Domain 1. Potential eligibility for available therapies or clinical trials (options: 0 / 5 / 10 points)

- **10** = Correct and comprehensive: all listed trial identifiers are relevant, with no incorrect entries and no omissions (within the prespecified reference list).
- **5** = Partially correct: trial identifiers are relevant and **contain no incorrect entries**, but there are **omissions** (i.e., missing some relevant trials) and/or the list is incomplete.
- **0** = Incorrect or misleading: includes any clearly incorrect trial identifier(s) for the genotype, or the recommendation is inconsistent with the prespecified evidence base.

Note: For genotypes with no approved therapy/clinical trial in our prespecified knowledge base, correctly stating "no relevant approved therapy/clinical trial" should be scored as 10 (complete and correct), not penalized.

Domain 2. Recommended follow-up interval (range: 0-10 points)

- **10** = Completely consistent with expert judgment and prespecified genotype-specific guidance (appropriate interval).
- **0** = Completely inconsistent with expert judgment or missing (e.g., clearly too long/too short given genotype-specific progression risk, or omitted).

Note: Intermediate scores (1-9) may be used when partially consistent.

Domain 3. Referral for genetic (reproductive) counseling (range: 0-10 points)

Item 3.1: Inheritance pattern (option: 0 or 3)

- **3** = Correct inheritance mode for the top-1 predicted genotype.
- **0** = Incorrect or missing.

Item 3.2: Partner-related counseling (range: 0-3)

- **3** = Fully consistent with expert judgment (e.g., appropriate recommendation regarding partner testing when indicated).
- **0** = Inconsistent with expert judgment or missing when indicated.

Item 3.3: Offspring risk counseling (range: 0-3)

- **3** = Fully consistent with expert judgment.

- **0** = Inconsistent with expert judgment or missing when indicated.

Item 3.4: Professionalism of the description (option: 0 or 1)

- **1** = Clear, clinically interpretable, and non-misleading.
- **0** = Ambiguous, misleading, or contains inappropriate wording.

Note: If the inheritance pattern is incorrectly identified, the score for this domain is 0.

Domain 4. Referral for syndromic/systemic evaluation (range: 0-10 points)

Item 4.1: Judgment on necessity (range: 0-8)

- **8** = Completely consistent with expert judgment and genotype-specific guidance (appropriate "refer" or "no referral needed").
- **0** = Completely inconsistent with expert judgment or missing (e.g., fails to recommend referral when clearly indicated, or recommends unnecessary referral).

Item 4.2: Specification of specialties/rationale (range: 0-2)

- **2** = Specialty selection and rationale are explicit, appropriate, and genotype-consistent (e.g., audiology for suspected Usher; nephrology/endocrinology for Bardet-Biedl; endocrine/cardiometabolic evaluation for Alström, when applicable).
- **0** = Ambiguous, misleading, or incorrect.

Note: If systemic referral is correctly judged as unnecessary, this domain can still score high.

Total score (range: 0-40): Sum of four domain scores; These recommendations were applied only to cases with a correct top-1 genotype prediction, as confirmed by NGS.

Supplementary Information - Section B

CLINICAL STUDY PROTOCOL

A Multi-stage Evaluation Foundation Model for Inherited Retinal Diseases Care Needs: A Randomized Control Trial

Protocol Number: EyeFM-IRD

Version Number: V1.2

Investigator Protocol Signature Page

PROTOCOL TITLE: A Multi-stage Evaluation Foundation Model for Inherited Retinal Diseases Care Needs: A Randomized Control Trial

I have carefully read this study protocol and agree that it contains all necessary information required to conduct this study. I agree to conduct the study according to this protocol (including any amendment) and in accordance with clinical sites Standard Operating Procedures, all other applicable regulations, and the recommendations laid down in the most recent version of the Declaration of Helsinki.

Signature of Principal Investigator	Date

Xiaodong Sun

Printed Name of Principal Investigator

Protocol Synopsis

Study Title	A Multi-stage Evaluation Foundation Model for Inherited Retinal Diseases Care Needs: A Randomized Control Trial
Protocol Number	EyeFM-IRD
Version Number	V1.2
Phase of Study	NA
Study Populations	Patients with suspected IRD, male or female
Study Objectives	To investigate the efficacy of Retina4IRD (EyeFM-IRD) on the diagnosis of the specific genetic mutation in suspected IRD patients.
Study Design	A multicenter, randomized, control trial.
Arms	The eligible patients will be randomly assigned in a 1:1 ratio to one of two arms. All enrolled patients meeting eligibility will be suggested to conduct the color fundus photographs (CFPs) and optical coherence tomography (OCT) images, as well as the NGS test, followed by: Arm A: AI assisted diagnostic arm (AI-based CDSS group) Arm B: specialist-only diagnostic arm without AI (control group)
Study Endpoints	**Primary Endpoint:** The top-5 gene mutation prediction accuracy, which will be assessed after all participants of the study have had next generation sequencing (NGS) test. **Secondary Endpoints:** The accuracy of top-1 to top-4 gene mutation predictions. All participants subsequently underwent confirmatory NGS testing, which served as the gold standard for genetic diagnosis. Blinded assessors compared the predictions from both groups against the NGS results to evaluate diagnostic accuracy.
Sample Size	A minimal sample size of 240 for statistical power above 95% with a two side Alpha = 0·05, estimating a top-5 gene mutation prediction accuracy of 85% in the AI-based CDSS group versus 60% in the specialist-only group. All randomly assigned cases were included in the primary analysis. All hypotheses tested were two-sided, and a *P* value of less than 0·05 was considered statistically significant.
Inclusion Criteria	**General Inclusion Criteria** Potential participants are eligible to be included in the study only if all of the following criteria apply: 1. Presenting with clinical features suggestive of suspected IRD based on the initial assessment by the physician.
Exclusion Criteria	**General exclusion criteria** Patients will be excluded from the study entry if any of the following criteria apply:

	1. Refusal to undergo NGS genetic testing, fundus photography, or OCT examination; 2. Presence of other ocular conditions that may interfere with retinal examination; 3. History of intraocular surgery in either eye within 6 months prior to screening; 4. Systemic severe diseases, intellectual disability, or psychiatric disorders; 5. Participation in other interventional clinical trials, such as stem cell or gene therapy trials.
Subject Discontinuation/ Withdrawal Criteria	Patients may withdraw from the study at his/her own request, or at the discretion of the Investigator at any time. The reasons for withdrawal include but are not limited to: • Withdrawal of consent. • Violation of the inclusion/exclusion criteria. • The subject has an AE and is unsuitable for the study as judged by the Investigator. • The subject experiences events which significantly affect safety, tolerability, and is unsuitable for the study as judged by the Investigator. • Patients have poor compliance and major protocol deviations occur. • The Investigator no longer believes participation is in the best interest of the participant.
Safety Evaluation	During the study period, any AEs that occur in all patients will be observed and recorded. The latest version of the Medical Dictionary for Regulatory Activities will be used to code adverse events by system organ class and preferred term.
Statistical methods	The AI model was evaluated via Top-N Accuracy. Analyses were performed using Python. The Top-N accuracy for each-class is defined as follows: $$Accuracy_c^N = \frac{M_c}{N_c}$$ Where $Accuracy_c^N$ is the Top-N accuracy for class c, N_c is the total number of samples in class c, M_c is the number of samples in class c for which the true label appears in the Top-N predictions. The overall Top-N accuracy is defined as follows: $$Accuracy_{overall}^N = \frac{\sum M_c}{\sum N_c}$$ Where $Accuracy_{coverall}^N$ is the Top-N accuracy over all samples, $\sum N_c$ is the total number of samples across all classes, $\sum M_c$ is the number of total samples for which the true label appears in the Top-N predictions.

Abbreviations

Abbreviations	Term
ADL	Activities of Daily Living
AE	Adverse Event
CDSS	Clinician-Decision Support System
CFP	Color Fundus Photographs
CI	Confidence Interval
DMP	Data Management Plan
DVP	Data Verification Plan
EDC	Electronic Data Capture
IRD	Inherited Retinal Diseases
NGS	Next Generation Sequencing
OCT	Optical Coherence Tomography
SAE	Serious Adverse Event

Table of Contents

1 Background Information

1.1 Disease and Target

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 cases worldwide.[2,3] To date, more than 300 nuclear and mitochondrial genes have been associated with IRDs, exhibiting immense phenotypic and genotypic heterogeneity. Precise genetic diagnosis is the critical first step for targeted treatment selection, prognostic outcomes evaluation, and genetic counseling, serving as the foundational gateway to personalized medicine for these conditions.

However, the diagnostic paradigms for IRDs face substantial and multifaceted challenges due to their heavy reliance on resource-intensive approaches. The current gold standard often requires multidisciplinary teams in specialized tertiary centers to integrate a complex array of data. This includes comprehensive ophthalmic evaluations (such as advanced imaging and electrophysiology), next-generation sequencing (NGS) often encompassing large multi-gene panels or whole-exome sequencing, meticulous familial segregation analysis, and in some cases, even functional validation of variants of uncertain significance.[3] This laborious and sequential workflow results in prolonged diagnostic odysseys that can last years, incur high costs for patients and healthcare systems, and offer limited geographical accessibility. Consequently, over 40% of patients with IRDs remain without a definitive genetic diagnosis despite extensive testing.[4,5] These systemic challenges are exacerbated by a critical global shortage of clinicians and genetic counselors skilled in interpreting genotypic-phenotypic correlations in retinal disorders. This expertise gap is particularly acute in resource-limited and rural settings, where misdiagnosis is common and diagnostic delays often exceed 5 years.[6] Such profound delays can have dire consequences, as they may cause patients with treatable conditions—such as those eligible for *RPE65* gene therapy or clinical trials—to miss critical therapeutic windows for advanced interventions.[7-9] This diagnostic bottleneck not only impedes clinical care and therapeutic advancement but also perpetuates significant psychological and emotional strain on affected families, creating an urgent and unmet need for more streamlined, accessible, and cost-effective diagnostic solutions.

1.2 Investigational AI model

EyeMF-IRD (also referred to as Retina4IRD), is engineered to tackle the complex challenge of IRD genotyping. The system is built upon a sophisticated Vision Transformer architecture, chosen for its superior ability to capture long-range dependencies and intricate patterns within medical images. To ensure robust feature extraction from the outset, we did not initialize our model from scratch. Instead, we leveraged a powerful visual foundation model that had been extensively pretrained on a massive, diverse dataset of over 0.9 million color fundus photographs (CFPs) and 0.7 million optical coherence tomography (OCT) images. This strategic pretraining step equips EyeMF-IRD with a rich, generalized understanding of retinal anatomy and pathology, providing a significant performance advantage and reducing the risk of overfitting on our specialized task.

For the specific objective of genotype diagnosis, we then meticulously trained and validated this pretrained model on a large-scale, multi-source, multimodal dataset. This comprehensive dataset was curated to reflect real-world clinical complexity, comprising 5,167 high-quality CFPs and OCT images and 8,901 structured clinical metadata items from 1,843 patients, ensuring diagnostic certainty for model training. To rigorously test the generalizability and mitigate geographic or center-specific biases, the data was collected through an international collaboration across China, South Korea, and Poland. This diverse geographic

origin strengthens the validation of our model and suggests its potential utility across different populations and healthcare settings.

1.3 Study Rationale

The EyeFM-IRD (also referred to as Retina4IRD) presented preliminary ideal performance in the diagnosis of the specific genetic mutations:

The performance of the EyeFM-IRD model was rigorously evaluated across internal and external validation sets, demonstrating its strong predictive capability and generalizability. When its analysis was based solely on CFP—a widely available and relatively inexpensive imaging modality—the model already achieved a robust top-5 prediction accuracy of 0.846 (95% CI: 0.842-0.850) on the internal dataset. This high performance was maintained on a completely unseen external dataset, where it attained an accuracy of 0.782 (95% CI: 0.776–0.788), confirming that the model had learned generalizable features rather than simply memorizing site-specific data.

Critically, the model's diagnostic power was significantly enhanced through a multimodal approach. The integration of OCT images, which provides detailed cross-sectional structural information of the retina, led to a substantial performance improvement. With this combined input, the top-5 prediction accuracy was 0.904 (95%CI 0.896-0.912) and 0.856 (95%CI 0.850-0.863) for internal and external validation, respectively. This significant leap in accuracy underscores the complementary value of OCT data, enabling the model to detect subtle genotypic clues that may not be apparent on fundus photography alone.

Besides, the model's utility was proven across diverse international populations. EyeFM-IRD maintained consistently robust performance when validated on separate cohorts from South Korea and Poland. This successful external validation across distinct geographic and ethnic backgrounds is a strong indicator of the model's robustness and its potential for global deployment, suggesting it can effectively assist in genetic diagnosis regardless of regional variations in patient demographics or clinical practices.

2 Study Objectives and Endpoints

2.1 Study Objectives

To investigate the efficacy of EyeFM-IRD on the diagnosis of the specific genetic mutation in suspected IRD patients.

2.2 Study Endpoints

2.2.1 Primary Endpoint:

The top-5 gene mutation prediction accuracy, which will be assessed after all participants of the study have had next generation sequencing (NGS) test

2.2.2 Secondary Endpoints:

The accuracy of top-1 to top-4 gene mutation predictions.

3 Study Design

3.1 Overall Study Design

This is a prospective, multicenter, randomized clinical trial to evaluate the efficacy of AI on the diagnosis of IRD.

The eligible patients will be randomly assigned in a 1:1 ratio to one of two arms. All enrolled patients meeting eligibility will be suggested to conduct the color fundus photographs (CFPs) and optical coherence tomography (OCT) images, as well as the NGS test, followed by:

Arm A: AI assisted diagnostic arm (AI-based CDSS group)

Arm B: specialist-only diagnostic arm without AI (control group)

3.2 Subject Withdrawal and Discontinuation of Study Intervention

3.2.1 Subject Withdrawal

Patients may withdraw from the study at his/her own request, or at the discretion of the Investigator at any time. The reasons for withdrawal include but are not limited to:

- Withdrawal of consent.
- Violation of the inclusion/exclusion criteria.
- The subject has an AE and is unsuitable for the study as judged by the Investigator.
- The subject experiences events which significantly affect safety, efficacy, and is unsuitable for the study as judged by the Investigator.
- Patients have poor compliance and major protocol deviations occur.
- The Investigator no longer believes participation is in the best interest of the participant.
- Decision by the sponsor to halt the entire study.

3.2.2 Discontinuation of Study Intervention

Study intervention discontinuation can be triggered by the participant (or legally authorized representative) or by the treating investigator.

Participants for whom study intervention is planned at any time during the study but not administrated will be considered to have temporarily discontinued study intervention. If study intervention is temporarily discontinued, it can be restarted at any time during the study. If study intervention is definitively discontinued, the participant will remain in the study to be evaluated for safety evaluation.

Patients who discontinue study treatment prematurely will be asked to return to the clinic for an early termination visit and may undergo assessments as outlined in the Schedule of Assessment. The primary reason for premature study treatment discontinuation should be documented on the appropriate eCRF.

3.3 Criteria for Premature Study Termination

Reasons for premature termination of the study or closing of the study site include but are not limited to:

- A significant unexpected AE occurs during the study, and the Investigator considers that the study should be terminated.
- The study shall be terminated due to major violations of the clinical study protocol, or relevant regulations during the study.

- The Sponsor requests termination under the premise of fully protecting the rights and safety of the subjects (for financial reasons, management reasons, etc.).
- The regulatory authority or Ethics Committee elects to terminate the study for certain reasons.

3.4 Definition of Study Completion

The study completion is defined as that the last subject completes his/her last visit, unless the discontinuation of the study is due to the Criteria for Premature Study Termination.

4 Study populations

4.1 Inclusion Criteria

General Inclusion Criteria

Potential participants are eligible to be included in the study only if all of the following criteria apply:

1. Presenting with clinical features suggestive of suspected IRD based on the initial assessment by the physician.

4.2 Exclusion Criteria

General exclusion criteria

Patients will be excluded from the study entry if any of the following criteria apply:

1. Refusal to undergo NGS genetic testing, fundus photography, or OCT examination;
2. Presence of other ocular conditions that may interfere with retinal examination;
3. History of intraocular surgery in either eye within 6 months prior to screening;
4. Systemic severe diseases, intellectual disability, or psychiatric disorders;
5. Participation in other interventional clinical trials, such as stem cell or gene therapy trials.

4.3 Screen Failures

Screen Failures are defined as subjects who signed the informed consent and do not meet the screening criteria during the screening period. Subjects who fail screening should not continue the study. The Investigator shall record the reasons for the screen failure in the original record.

4.4 Subject Management During the Study

4.4.1 Concomitant Medications

Any medication considered necessary for the subject's welfare, and that are not expected to interfere with the evaluation of the study drug, may be given at the discretion of the investigator, with the exceptions noted below:

Subjects may not receive any standard or investigational agents for treatment of their IRD, such as stem cell or gene therapy trials.

4.4.2 Relative Examination

For Arm A & B, if any discomfort occurs when the examinations were conducted, additional measurement will be permitted based on investigator discretion.

4.4.3 Subject Compliance Management

Before the study, Investigators should ensure that patients are fully informed and understand the IMP and study procedures and increase the confidence of the patients in completing the study. The patients should also be told in detail about the precautions and possible adverse reactions.

5 Study Intervention

5.1 Investigational Model

EyeFM-IRD, a multimodal foundation model-based clinician-decision support system (CDSS) that replicates a clinician's diagnostic reasoning pathway, through integrated analysis of retinal imaging and key clinical metadata.

5.2 Randomization and Masking

The study employed a computer-generated pseudo-random number sequence to randomize participants in a 1:1 ratio. After eligibility assessment, all enrolled subjects were automatically assigned through the study platform to either the Retina4IRD-assisted diagnostic arm (AI-based CDSS group) or the specialist-only diagnostic arm (control group).

Participants in the AI-based CDSS group underwent basic medical history documentation, fundus photography, and OCT imaging. These data were then processed by the Retina4IRD system, which generated a list of the top five predicted genetic mutations for physician review and final diagnosis. In the control group, participants received the same baseline assessments. However, physicians in this arm independently evaluated whether patients had an IRD and provided their own top five genetic mutation predictions without AI assistance, relying solely on the available CFP, OCT, and clinical metadata.

All participants subsequently underwent confirmatory NGS testing, which served as the gold standard for genetic diagnosis. Blinded assessors compared the predictions from both groups against the NGS results to evaluate diagnostic accuracy.

5.3 Administration

Patients will be given normative diagnosis of IRD based on CFP, OCT, and clinical metadata items, and further confirmed via NGS test.

All investigators, regardless in arm A or arm B, were qualified and trained in administering patients with suspected IRD, and followed standard diagnosis guidelines.

6 Study Procedures

This study is conducted as a diagnostic test, aimed to evaluate the performance of Retina4IRD in real-world clinical setting. All study assessments are as described below and in **Section 7** and **8**.

6.1 Consent and Screening

- The suspected IRD patients with OCT and CFP examinations would be recruited;
- Voluntary, written informed consent using ICF is required prior to any screening or study procedure as per **Section 11.3**;
- Record demography;
- Record patient's family history and the onset age of the IRD;

6.2 Diagnosis of the genetic mutations

Based on the CFP, OCT, and clinical metadata, the physicians need to diagnose as below:

Arm A: In the diagnostic process of IRD, retinal specialists make diagnostic decision with the assistance of FM-IRDs;

Arm B: In the diagnostic process of IRD, retinal specialists make independent decision-making without the assistance of FM-IRDs;

Finally, the NGS testing would provide the gold standard.

7 Efficacy Evaluation

7.1 Top-5 gene mutation prediction accuracy

The primary outcome measure of the RCT is the top-5 gene mutation prediction accuracy, which will be assessed after all participants of the study have had NGS test.

7.2 Top-1 to Top-4 gene mutation prediction accuracy

The secondary endpoint is the accuracy of top-1 to top-4 gene mutation predictions. The Top-N accuracy for each class is defined as follows:

$$Accuracy_c^N = \frac{M_c}{N_c}$$

Where $Accuracy_c^N$ is the Top-N accuracy for class c, N_c is the total number of samples in class c, M_c is the number of samples in class c for which the true label appears in the model's Top-N prediction.

8 Safety Evaluation

8.1 Safety Evaluations

Safety evaluations include incidence and severity of ocular.

8.2 Adverse Events

The investigator is responsible for detecting and recording adverse events (AEs) and serious adverse events (SAEs) that meet the definitions in this protocol.

8.2.1 Definitions

Adverse Events

An AE is any untoward medical occurrence in a patient or clinical investigation subject which does not necessarily have a causal relationship with the intervention. An AE can therefore be any unfavorable sign, symptom, or disease temporally associated with the intervention, whether or not related to the intervention.

Serious Adverse Events

Serious Adverse Event (SAE): Any untoward medical occurrence that at any arm results in death, is life-threatening, requires inpatient hospitalization or prolongation of existing hospitalization, results in persistent or significant disability/incapacity, or is a congenital anomaly/birth defect.

A SAE is any AE that meets any of the following criteria:

- Death: When the outcome of the medical event is death, it can be recorded and reported as a SAE definitely;
- Life-threatening: an event in which the subject was at risk of death at the time of the event, which hypothetically might have caused death if it is more severe;
- Requires inpatient hospitalization or prolongation of existing hospitalization: it is necessary to make it clear that it occurs due to the AE, not due to elective surgery, non-medical reasons, etc.;
- Results in persistent or significant disability/incapacity;
- Congenital anomaly/birth defect in a neonate/infant;
- Other important medical events: Important medical events that may not result in death, be life-threatening, or require hospitalization may be considered serious when, based upon appropriate medical judgment, they may jeopardize the subject and may require medical or surgical intervention to prevent one of the outcomes listed in the definition above.

8.2.2 Severity of AEs

Investigators will classify the severity of AEs into 5 grades:

- Grade 1: Mild; asymptomatic or mild symptoms; clinical or diagnostic observations only; intervention not indicated.
- Grade 2: Moderate; minimal, local or noninvasive intervention indicated; limiting age-appropriate instrumental activities of daily living (ADL). ADLs include cooking, shopping groceries or clothes, using the phone, managing finances, etc.
- Grade 3: Severe or medically significant but not immediately life-threatening; hospitalization or prolongation of hospitalization indicated; disabling; limiting self-care ADL. The self-care ADL includes bathing, getting dressed, eating, toileting, taking

medicine, not bedridden.

- Grade 4: Life-threatening consequences; urgent intervention indicated.
- Grade 5: Death related to the AE.

8.2.3 SAEs Reactions Reporting

T If a serious adverse event occurs during the trial, the investigator shall immediately take necessary measures to protect the safety of the subjects. And report it to the Ethics Committee within 24 hours.

9 Statistical Analysis Plan

9.1 Sample Size

Statistical hypothesis: the following null hypothesis H_0 will be tested against the alternative hypothesis, H_1, to assess optimal efficiency for primary endpoint:

- H_0: performance of arm A in Top-5 genetic mutation prediction accuracy is not more effective than the counterpart of arm B;
- H_1: performance of arm A in Top-5 genetic mutation prediction accuracy is more effective than the counterpart of arm B.

The **primary endpoint** is considered successful if the null hypothesis H_0 is rejected, calculated via the z-test using the top-5 accuracy for the two arms.

In the planning stages of this randomized controlled trial, a formal sample size calculation is conducted to ensure the study is sufficiently powered to detect a meaningful difference between the two diagnostic arms. The calculation is based on the primary outcome of top-5 gene mutation prediction accuracy. We hypothesize an accuracy of **85%** in the AI-based clinical decision support system group, compared to **60%** in the specialist-only control group. To detect this anticipated 25-percentage-point absolute difference with a statistical power of above **95%** (i.e., a less than 5% chance of a Type II error) and a two-sided alpha level of **0.05** indicating a 5% risk of a Type I error), a minimum sample size of 240 participants is required. This ensured the trial is rigorously designed to identify a significant effect of the intervention.

The primary efficacy analysis was pre-specified to be performed. This conservative approach analyzes all randomly assigned participants in the groups to which they were originally allocated, regardless of protocol deviations, withdrawal, or the actual use of the Retina4IRD system. By including every case in the analysis, the principle preserves the baseline randomization, minimizes bias, and provides a pragmatic estimate of the intervention's effectiveness in a real-world clinical setting. All statistical hypotheses tested in this trial were two-sided, and a *P* value of less than 0.05 was considered the threshold for statistical significance.

9.2 Endpoints and Analysis

Primary Endpoint: The top-5 gene prediction accuracy, which will be assessed after all participants of the study have had next generation sequencing (NGS) test.

Secondary Endpoints: The accuracy of top-1 to top-4 gene predictions.

The Top-N accuracy for each-class is defined as follows:

$$Accuracy_c^N = \frac{M_c}{N_c}$$

Where $Accuracy_c^N$ is the Top-N accuracy for class c, N_c is the total number of samples in class c, M_c is the number of samples in class c for which the true label appears in the Top-N predictions. The overall Top-N accuracy is defined as follows:

$$Accuracy_{overall}^N = \frac{\sum M_c}{\sum N_c}$$

Where $Accuracy_{coverall}^N$ is the Top-N accuracy over all samples, $\sum N_c$ is the total number of samples across all classes, $\sum M_c$ is the number of total samples for which the true label appears in the Top-N predictions.

Exploratory Endpoint: In real-world IRD care, diagnostic confirmation is the first step, four questions are typically asked by patients and are currently provided by clinicians. (1) potential eligibility for approved new therapies or trials; (2) follow-up intervals based on genotype-specific natural history; (3) need for genetic/reproductive counseling; and (4) need for multidisciplinary referral, given syndromic associations. Accordingly, we developed a composite Downstream Management Strategy Score to post hoc evaluate Retina4IRD's utility for downstream clinical management.

The composite strategy score covering four dimensions: (i) potential eligibility for available therapies or clinical trials, (ii) recommended follow-up interval, (iii) referral for genetic (reproductive) counseling, and (iv) referral for syndromic/systemic evaluation. Specifically, we provided the AI with knowledge of clinical and genetic information for IRD mutation 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; clinicians then complete the management report after reviewing the AI output. In the control arm, clinicians 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, 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.

Independent two-proportion z-test was used to compare Top-N accuracy between the AI-assisted and control groups. A two-sample t-test was performed to compare the composite scores and dimension-specific scores between the two groups. All hypothesis tests were two-sided, and a *P* value of less than 0.05 was considered statistically significant. Statistical analyses were conducted using Python.

9.3 Demographics and Baseline Characteristics

Baseline characteristics will be presented in listings and summarized using descriptive statistics. All study variables will be described via median [minimum, maximum] or n (%), as appropriate.

9.4 Report of Adverse Events

Ocular adverse events and non-ocular adverse events will be summarized separately. Ocular adverse events will be also classified as either study eye or non-study eye (fellow eye) events. In addition, listings of AEs leading to discontinuation of the study, SAEs and deaths, will be provided as applicable.

10 Quality Control and Quality Assurance

The investigator is responsible for formulating, implementing and timely updating the standard operating procedures related to the clinical study quality assurance and quality control system, ensuring that the implementation of the clinical study, data generation, records and reports all comply with the study protocol and other requirements of relevant laws and regulations.

The entire process of the clinical study must be carried out in strict accordance with the standard operating procedures for quality management.

Quality control should be performed in each stage of data processing to ensure that all data is reliable and the data processing procedure is correct.

11 Ethics

11.1 Statement

The implementation of this clinical study strictly follows the Declaration of Helsinki (2013 edition), ICH E6 (R2) (2016 edition), and other relevant laws and regulations.

11.2 Ethics Committee

The responsibility of the Ethics Committee is to protect the rights and interests and safety of subjects, and special attention should be paid to vulnerable subjects.

The Ethics Committee should review the scientific issues and ethics of the clinical study and the qualifications of Investigators. The documents reviewed include: the study protocol and the revised version of the study protocol, the ICF and its updates. The Ethics Committee should complete the review or filing process of clinical study-related materials within a reasonable time limit and give clear written review opinions.

Before the start of the clinical study, the approval of the Ethics Committee must be obtained before this study can be started. During the clinical study, any modification to the study protocol must be approved or filed by the Ethics Committee, and the Ethics Committee should be notified after the study is over.

The Ethics Committee should pay attention to and clearly require Investigators to report in a timely manner: deviation from or modification to the study protocol during the implementation of the clinical study to eliminate emergency hazards to subjects; changes that increase the risk of subjects or significantly affect the implementation of the clinical study; all suspected unexpected serious adverse reactions; new information that may adversely affect the safety of subjects or the implementation of the clinical study.

11.3 Informed Consent Form

Eligible participants may only be included in the study after providing voluntary, written (witnessed, where required by law or regulation), IRB/IEC-approved informed consent. Informed consent to participate in the study must be obtained according to local regulatory requirements before conducting any study-specific procedures (i.e., all of the procedures described in the protocol).

If new information becomes available that may be relevant to the participant's willingness to continue to participate in the study, a new informed consent form will be prepared and approved by the EC. The study participants will be informed about the new information and re-consent obtained. Participants who are re-screened must sign a new ICF.

11.4 Subject Privacy Protection

The privacy of subjects and the confidentiality of related information should be protected.

A unique code will be assigned to each subject to identify their identity in the clinical study, that is, subject identification code. When the Investigator reports the AEs and other study-related data of the subject, the code will be used to replace the subject's name to protect their privacy.

When conducting direct access, any party who has such direct access should take reasonable measures to protect the privacy of subjects in accordance with relevant laws and regulations. In the process of processing clinical study information and subject information, care should be taken to avoid illegal or unauthorized access, disclosure, dissemination, modification, damage, and loss of information. The recording, processing and preservation of clinical study data should ensure the confidentiality of records and subject information.

12 Protocol Modification and Deviation

The Investigator should implement the clinical study in accordance with the study protocol approved by the Ethics Committee.

Without the consent of the Ethics Committee, the Investigator may not modify or deviate from the study protocol, excluding changes made to eliminate emergency hazards to the subjects or change of telephone numbers, and changes alike that only involve the management of the clinical study.

The Investigator or his/her designee should record and explain any deviation from the study protocol.

If the Investigator modifies or deviates from the study protocol without the approval of the Ethics Committee in order to eliminate emergency hazards to the subjects, he/she should report to the Ethics Committee in a timely manner, explain the reasons.

13 Reference

1. Georgiou M, Robson AG, Fujinami K, 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* 2024; **100:** 101244.
2. Ben-Yosef T. Inherited Retinal Diseases. *Int J Mol Sci* 2022; **23:** 13467.
3. Schneider N, Sundaresan Y, Gopalakrishnan P, et al. Inherited retinal diseases: Linking genes, disease-causing variants, and relevant therapeutic modalities. *Prog Retin Eye Res* 2022; **89:** 101029.
4. Britten-Jones AC, Gocuk SA, Goh KL, Huq A, Edwards TL, Ayton LN. The Diagnostic Yield of Next Generation Sequencing in Inherited Retinal Diseases: A Systematic Review and Meta-analysis. *Am J Ophthalmol* 2023; **249:** 57–73.
5. Sheck LHN, Esposti SD, Mahroo OA, et al. Panel-based genetic testing for inherited retinal disease screening 176 genes. *Mol Genet Genomic Med* 2021; **9:** e1663.
6. Nguyen Q, Woof W, Kabiri N, et al. Can artificial intelligence accelerate the diagnosis of inherited retinal diseases? Protocol for a data-only retrospective cohort study (Eye2Gene). *BMJ Open* 2023; **13:** e071043.
7. Pierce EA, Aleman TS, Jayasundera KT, et al. Gene Editing for *CEP290*-Associated Retinal Degeneration. *N Engl J Med* 2024; **390:** 1972–84.
8. Botto C, Rucli M, Tekinsoy MD, Pulman J, Sahel JA, Dalkara D. Early and late stage gene therapy interventions for inherited retinal degenerations. *Prog Retin Eye Res* 2022; **86:** 100975.
9. Suh S, Choi EH, Raguram A, Liu DR, Palczewski K. Precision genome editing in the eye. *Proc Natl Acad Sci U S A* 2022; **119:** e2210104119.

CONSORT 2010 checklist of information to include when reporting a randomised trial

Section/Topic	Item No	Checklist item	Reported on page No
Title and abstract			
	1a	Identification as a randomised trial in the title	Page 1
	1b	Structured summary of trial design, methods, results, and conclusions (for specific guidance see CONSORT for abstracts)	Page 4
Introduction			
Background and objectives	2a	Scientific background and explanation of rationale	Page 5
	2b	Specific objectives or hypotheses	Page 6
Methods			
Trial design	3a	Description of trial design (such as parallel, factorial) including allocation ratio	Page 26 & 30-31
	3b	Important changes to methods after trial commencement (such as eligibility criteria), with reasons	NA
Participants	4a	Eligibility criteria for participants	Page 30-31
	4b	Settings and locations where the data were collected	Page 30
Interventions	5	The interventions for each group with sufficient details to allow replication, including how and when they were actually administered	Page 31
Outcomes	6a	Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed	Page 32
	6b	Any changes to trial outcomes after the trial commenced, with reasons	NA
Sample size	7a	How sample size was determined	Page 33-34
	7b	When applicable, explanation of any interim analyses and stopping guidelines	NA
Randomisation:			
Sequence generation	8a	Method used to generate the random allocation sequence	Page 31
	8b	Type of randomisation; details of any restriction (such as blocking and block size)	Page 31
Allocation concealment mechanism	9	Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned	Page 31
Implementation	10	Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions	Page 31
Blinding	11a	If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how	Page 31
	11b	If relevant, description of the similarity of interventions	Page 31-32
Statistical methods	12a	Statistical methods used to compare groups for primary and secondary outcomes	Page 33-34
	12b	Methods for additional analyses, such as subgroup analyses and adjusted analyses	Page 34
Results			
Participant flow (a diagram is strongly recommended)	13a	For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analysed for the primary outcome	Figure 4
	13b	For each group, losses and exclusions after randomisation, together with reasons	Figure 4
Recruitment	14a	Dates defining the periods of recruitment and follow-up	Page 9
	14b	Why the trial ended or was stopped	NA
Baseline data	15	A table showing baseline demographic and clinical characteristics for each group	Table 1

Section/Topic	Item No	Checklist item	Reported on page No
Numbers analysed	16	For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups	Table 1
Outcomes and estimation	17a	For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)	Page 9-10
	17b	For binary outcomes, presentation of both absolute and relative effect sizes is recommended	Page 9-10
Ancillary analyses	18	Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory	Page 10
Harms	19	All important harms or unintended effects in each group (for specific guidance see CONSORT for harms)	Page 10
Discussion			
Limitations	20	Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses	Page 13-14
Generalisability	21	Generalisability (external validity, applicability) of the trial findings	Page 13
Interpretation	22	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence	Page 13
Other information			
Registration	23	Registration number and name of trial registry	Page 4
Protocol	24	Where the full trial protocol can be accessed, if available	Supplementary information - Section B
Funding	25	Sources of funding and other support (such as supply of drugs), role of funders	Page 14