Benchmarking Foundation Models for Early Multi-Retinopathy Detection at National Scale: Performance, Generalization, and Bias

Wenquan Cheng[1], Jia Guo[1], Jinyuan Wang[2], Yihua Sun[1], Haojie Han[1], Fang Chen[3], Hongen Liao[1,3], Tien Yin Wong[1,4,5], Xiaodong Sun[6], and Huixun Jia[6]

[1] School of Biomedical Engineering, Tsinghua Medicine, Tsinghua University, Beijing, China
wongtienyin@tsinghua.edu.cn
[2] School of Clinical Medicine, Tsinghua Medicine, Tsinghua University, Beijing, China
[3] School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
[4] Beijing Visual Science and Translational Eye Research Institute (BERI), Tsinghua Medicine, Tsinghua University, Beijing, China
[5] Singapore Eye Research Institute, Singapore National Eye Centre, Singapore
[6] Department of Ophthalmology, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China
xdsun@sjtu.edu.cn

Abstract. Vision impairment affects 2.2 billion people globally, yet scalable screening is hindered by workforce shortages. We introduce OphFM-bench, a national-scale benchmark for early multi-retinopathy screening using color fundus photographs. The dataset includes over half a million adults from 1,010 hospitals across 27 Chinese provinces, capturing diverse devices and population structures. We evaluate three foundation model families—retina-specific models, general medical vision-language models (VLMs), and general-purpose backbones across three dimensions: Screening Efficiency (high-sensitivity regimes), Spatial Generalization (geographical shifts), and Algorithmic Fairness (age, sex, and residency). While retina-specific models like RETFound-DINO excel in ranking, medical VLMs like MedSigLIP offer superior transfer efficiency and robustness in high-recall scenarios. Notably, we identify a performance-calibration gap where high AUC masks significant false-positive burdens in low-prevalence areas. OphFM-bench establishes an actionable framework for equitable AI deployment in national-scale screening networks.

Keywords: Real-world Screening · Foundation Model · Ophthalmology.

1 Introduction

Retinal diseases such as diabetic retinopathy (DR), age-related macular degeneration (AMD), pathological myopia (PM), and retinal vein occlusion (RVO) are

X. Sun and T. Y. Wong are the co-corresponding authors.

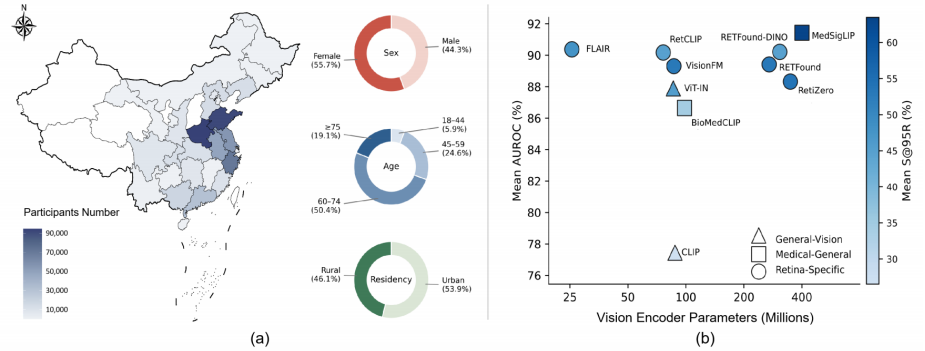

Fig. 1. Cohort characterization and model scaling. (a) Geographic and demographic distribution of participants across Chinese provinces. (b) Performance scaling with model size.

leading causes of preventable vision impairment [6,8]. Early detection and timely intervention can avert irreversible vision loss, yet population-level screening is difficult to scale due to specialist shortages, geographic inequities, and variable grading consistency across institutions [7, 18].

Teleophthalmology partially alleviates these bottlenecks by enabling peripheral acquisition of fundus images with centralized triage [17, 20]. However, real-world deployment remains challenged by heterogeneous devices and workflows, severe class imbalance, and label uncertainty [2].

Meanwhile, foundation models (FMs) have shifted visual representation learning from single-task supervised training to scalable self-supervised and vision-language pretraining. In ophthalmology, retina-specific FMs such as RETFound demonstrate strong transfer across retinal disease tasks [24]. In parallel, medical VLMs offer data-efficient adaptation and promise more robust transfer [22], but their screening behavior under real-world distribution shift is not well characterized [13].

Despite rapid progress, evidence for ophthalmic FMs is still limited by two fundamental gaps.(1) *The Dataset Gap:* There is a profound discrepancy between the homogenous, cleanly annotated public datasets or single-center retrospective cohorts traditionally utilized for model training, and the highly heterogeneous data derived from national natural populations [5]. Current datasets fail to capture real-world complexities, including epidemiological heterogeneity, diverse imaging devices and clinical workflows, severe label noise, and extreme class imbalance [9,12]. (2) *The Evaluation Gap:* The community lacks a unified, real-world evaluation framework tailored for early disease screening. Existing literature rarely moves beyond standard independent and identically distributed (i.i.d.) performance metrics, failing to simultaneously interrogate performance, generalization, and socioeconomic bias [23].

To address these gaps, we introduce **OphFM-bench**, a systematic benchmark for early multi-retinopathy screening from fundus photography at a na-

tional scale. We evaluate retina-specific FMs, medical VLMs, and general backbones across linear probing and fine-tuning, using a unified protocol that couples: (1) screening-oriented operating points, (2) spatial generalization across major geographic and socioeconomic shifts, and (3) fairness audits across age, sex, and urban-rural residency. Our goal is not merely to rank models, but to reveal where failures occur, identify adaptation regimes that remain reliable under national-scale shifts, and provide actionable guidance for equitable deployment in telemedicine screening networks.

2 Benchmark Construction

2.1 The National Telemedicine Cohort

We conduct a retrospective benchmark designed to reflect the full heterogeneity of routine tele-ophthalmology screening, rather than curated research cohorts. To avoid inflating performance via restrictive filtering, we apply no additional inclusion/exclusion criteria beyond data availability, preserving real-world variation in image quality, devices, workflows, and patient mix.

Coverage and clinical diversity. Data are aggregated from a distributed network of 1,010 medical centers across 129 cities in 27 provinces in China (Fig. 1a), spanning the full spectrum of care settings (top-tier hospitals, secondary hospitals, and primary/community facilities).

Population-scale class imbalance. The cohort preserves the natural prevalence of retinal diseases in a health-check population. Consequently, normal fundus images constitute the majority class with sparse disease labels, mirroring the statistical environment that determines real-world referral burden and operational feasibility in screening programs.

Participants and metadata. The dataset contains 503,689 anonymized adults undergoing routine health checks with color fundus photography, totaling 605,224 images. Each participant is associated with standardized questionnaire and examination metadata, including demographics, residential attributes, and medical histories. Urban–rural residency is derived from medical insurance information. All participants provided written informed consent and the Institutional Review Board of National Clinical Research Center for Ophthalmic Diseases approved the study protocol (2022KY-120). This study was conducted in accordance with Declaration of Helsinki.

Reference labels via telemedicine grading. Images are graded by expert ophthalmologists through a centralized telemedicine platform for signs of DR, AMD, PM, and RVO, yielding disease-specific labels for multi-disease screening evaluation. This setup reflects realistic label uncertainty and inter-rater variability expected in operational telemedicine pipelines.

2.2 Foundation Models Spanning Pretraining Domains

To evaluate the practical transferability of foundation models under real-world screening constraints, we benchmark models representing distinct pretraining

domains and learning objectives: (1)*Retina-Specific Foundation Models:* Architectures explicitly pre-trained on massive corpora of ophthalmic imaging data, designed to capture the unique topological and pathological features of the human retina (RETFound, RETFound-DINO [24], VisionFM [10], FLAIR [15], RetCLIP [4], and RetiZero [19]). (2)*Medical-General Foundation Models:* Models trained on diverse multi-modal medical datasets spanning radiology, pathology, and dermatology to establish generalized medical intelligence (MedSigLIP [14] and BiomedCLIP [22]).(3)*Domain-Agnostic General-Vision Models:* Baseline architectures representing state-of-the-art computer vision trained on massive natural images (DINOv3 [16], ViT-IN [3], and CLIP [11]).

2.3 Evaluation Framework for Screening

Performance and Transfer Efficiency. We quantify representation quality and data efficiency through two canonical adaptation regimes: linear probing (LP) and full fine-tuning (FT). In deployment, LP is attractive for efficiency and stability, while FT is often necessary when the task shift is large. Comparing both regimes provides a direct measure of transfer efficiency. The dataset is randomly divided into train/valid/test sets at a ratio of 5:1:24 at patient level, resulting in 96,883/24,190/484,151 images.

Model performance is quantified using the area under the receiver operating characteristic curve (AUC). However, screening pipelines require high recall to minimize missed diseases and high specificity to control downstream specialist workloads. Therefore, we extend our metrics beyond AUC to include clinically relevant operating points: Specificity at 95% Recall (S@95R), Specificity at 99% Recall (S@99R), and Recall at 95% Specificity (R@95S). This approach aligns with real-world regulatory endpoints for autonomous diagnostic systems [21].

Real-world Generalization. Screening algorithms must not silently degrade when deployed outside the training environment. We therefore evaluate robustness under a hard geographic distribution shift.

China provides a naturally occurring stress test for real-world deployment. The widely recognized North–South shift marks substantial differences in climate, ecology, and socioeconomic development. Beyond environment and healthcare access, large-scale genetics studies report measurable north–south population structure correlated with geography, which can co-vary with disease prevalence and comorbidity profiles [1]. We fine-tune models using South-only data in the training set and evaluate them on unseen North-only data in the test set. This setup directly tests whether a model remains reliable when facing multifactor shifts (population mix, devices, practice patterns) that resemble realistic telemedicine expansion across regions [21].

Fairness and Socioeconomic Bias. To support equitable deployment, we conduct subgroup analyses across three primary axes: sex, age, and urban versus rural residency. The continuous age variable is stratified into four clinically

Table 1. Performance metrics of FT (%). Best and second best per row are **bold** and underlined. Model abbreviations: FLR (FLAIR), RC (RetCLIP), RF (RETFound), RF-D (RETFound-DINO), RZ (RetiZero), VFM (VisionFM), BMC (BioMedCLIP), MSL (MedSigLIP), CLP (CLIP).

		Retina-Specific						**Medical**		**General**	
Disease	Metric	FLR	RC	RF	RF-D	RZ	VFM	BMC	MSL	CLP	ViT
AMD	AUC	86.73	86.66	86.11	**88.20**	85.28	84.61	83.52	87.75	69.03	84.84
	S@95R	49.84	46.40	44.98	54.99	40.67	39.59	33.83	**58.29**	17.24	40.83
	S@99R	25.91	27.20	28.39	**31.94**	23.73	24.10	23.11	28.54	15.83	24.40
	R@95S	42.12	42.24	44.19	**46.14**	42.87	42.08	40.57	42.90	20.54	41.60
DR	AUC	88.91	87.91	86.76	**89.74**	85.95	86.28	84.22	89.44	74.88	85.40
	S@95R	42.35	36.80	37.60	46.06	35.58	34.98	29.62	**46.82**	21.07	35.32
	S@99R	21.08	18.81	23.32	**27.70**	20.27	22.41	19.30	23.47	15.00	22.21
	R@95S	58.54	58.19	53.27	**61.77**	52.38	52.95	48.48	59.62	28.05	50.56
PM	AUC	94.67	94.25	95.19	95.12	95.39	95.49	93.86	**95.55**	92.39	94.70
	S@95R	64.12	51.23	50.16	65.08	65.11	67.44	45.04	**71.78**	41.48	54.60
	S@99R	31.09	18.92	18.02	42.40	36.88	35.99	21.36	**60.09**	11.66	12.15
	R@95S	77.57	79.04	**80.18**	80.02	79.56	78.60	76.34	79.05	69.87	76.64
RVO	AUC	91.17	91.87	89.58	87.77	86.66	90.80	84.95	**92.06**	73.45	86.59
	S@95R	34.80	44.42	**79.99**	13.11	73.68	70.27	28.84	77.04	26.23	51.71
	S@99R	27.03	29.13	**76.36**	11.93	31.74	36.57	22.37	47.12	18.72	11.86
	R@95S	75.48	**81.72**	75.05	74.62	62.37	75.48	60.43	80.43	32.04	64.95
Mean	AUC	90.37	90.17	89.41	90.21	88.32	89.30	86.64	**91.20**	77.44	87.88
	S@95R	47.78	44.71	53.18	44.81	53.76	53.07	34.33	**63.48**	26.51	45.62
	S@99R	26.28	23.52	36.52	28.49	28.16	29.77	21.54	**39.81**	15.30	17.66
	R@95S	63.43	65.30	63.17	**65.64**	59.30	62.28	56.46	65.50	37.63	58.44

meaningful cohorts ($\leq$ 44, 45-59, 60-74, and $\geq$ 75 years) to align with established epidemiological milestones in retinal pathology. Algorithmic fairness is mathematically evaluated for both LP and FT adaptation regimes using standard parity metrics, including AUC Disparity (ΔAUC) and equal opportunity (EOpp).

3 Experiments and Analyses

3.1 Performance and Transfer Efficiency

FT on test set. Across four diseases, retina-specific self-supervised models dominate AUROC on AMD and DR, while medical VLMs dominate screening-tail efficiency at high recall (Table 1). This suggests that medically tuned image–text alignment produces features that remain robust when pushed to high-sensitivity regimes. A key takeaway is that AUROC ranking and screening efficiency are not interchangeable: models with near-tied AUROC can differ sharply in tail specificity, which directly translates to referral burden.

Limitations in Autonomous Screening. Our evaluation reveals a critical performance asymmetry: $R@95S$ consistently outperforms $S@95R$ across all

Table 2. Performance metrics of LP (%). Best and second best per row are **bold** and underlined.

		Retina-Specific						Medical		General	
Disease	Metric	FLR	RC	RF	RF-D	RZ	VFM	BMC	MSL	CLP	ViT
AMD	AUC	78.45	84.49	73.51	83.77	78.89	79.97	67.76	**85.85**	57.46	78.01
	S@95R	21.65	34.52	22.16	34.97	26.39	27.25	14.15	**47.20**	9.19	29.09
	S@99R	6.65	9.17	7.50	12.00	8.95	9.34	4.25	**18.49**	1.60	9.94
	R@95S	29.57	**41.08**	24.91	39.19	33.93	36.70	21.55	40.33	8.29	28.06
DR	AUC	83.42	85.57	79.29	84.82	81.73	82.06	72.58	**87.82**	64.68	81.61
	S@95R	22.99	30.65	22.49	30.36	21.44	24.00	15.08	**37.33**	10.67	26.80
	S@99R	6.87	9.15	5.84	7.58	6.34	5.36	3.63	**11.27**	1.60	8.75
	R@95S	52.13	52.34	37.31	52.02	44.43	47.13	29.21	**56.49**	12.52	40.05
PM	AUC	87.25	94.39	94.71	93.11	94.87	**95.19**	92.36	94.42	77.29	92.37
	S@95R	33.08	60.36	63.83	39.69	**65.69**	62.52	60.93	61.90	14.82	51.10
	S@99R	9.45	17.43	38.03	21.91	**43.16**	37.49	32.27	24.98	1.63	28.04
	R@95S	57.26	**79.93**	75.97	74.79	76.74	78.01	68.91	77.99	34.53	72.24
RVO	AUC	82.91	**92.36**	87.50	87.25	87.43	86.72	77.74	90.41	54.27	85.53
	S@95R	38.26	21.55	29.28	64.00	42.58	58.09	48.26	**72.41**	15.55	41.71
	S@99R	26.08	9.09	18.80	23.38	12.31	50.96	19.03	**56.38**	2.34	38.72
	R@95S	56.34	**80.22**	64.09	69.25	68.39	67.31	40.86	77.63	8.60	55.48
Mean	AUC	83.01	89.20	83.75	87.24	85.73	85.99	77.61	**89.63**	63.43	84.38
	S@95R	29.00	36.77	34.44	42.26	39.03	42.97	34.61	**54.71**	12.56	37.18
	S@99R	12.26	11.21	17.54	16.22	17.69	25.79	14.80	**27.78**	1.79	21.36
	R@95S	48.83	**63.39**	50.57	58.81	55.87	57.29	40.13	63.11	15.99	48.96

models. In populational screening, characterized by low disease prevalence, clinical and regulatory safety mandates necessitate operating at the $S@95R$ threshold to minimize life-altering false negatives. However, the observed suboptimal $S@95R$ implies that enforcing high recall triggers an unsustainable volume of false positives. In low-prevalence cohorts, this low specificity leads to a poor positive predictive value of unnecessary referrals. These results suggest that while current FMs excel in diagnostic assistance, they require significant recalibration or alignment to meet the rigorous specificity requirements for high-throughput, autonomous screening.

LP for transfer efficiency. LP isolates the quality of pretrained representations by training only a linear head, a standard probe for feature linear separability. MedSigLIP is the strongest LP model for AMD/DR, achieving the best AUC and best tail-specificities. This suggests that medical image–text pretraining yields features that are immediately separable for lesion-driven diseases, enabling highly data-efficient deployment. For RVO, RetCLIP achieves the best LP AUC, implying that ranking quality and extreme-tail specificity can be decoupled even under LP.

FT–LP performance shifts. Comparing LP to FT clarifies which models need task-specific adaptation to become screening-ready.MedSigLIP improves only modestly in AUC from LP to FT, but improves substantially in screening-tail specificity. This indicates its representations are already strong, while FT pri-

Table 3. Geographical generalization performance and cross-region retention (%). Each entry is formatted as *Target Region* (Retention,%), where Retention is $\frac{Target\ Region}{Source\ Region}$ and *Source Region* is the local fine-tuning result from Table 1. Best and second best results are **bold** and underlined.

Disease	Metric	FLR	RC	RF-D	RZ	MSL
AMD	AUC	85.21 (98.2)	85.13 (98.2)	**87.05** (98.7)	83.85 (98.3)	86.52 (98.6)
	S@95R	45.34 (91.0)	41.90 (90.3)	50.89 (92.5)	36.17 (88.9)	**54.19** (93.0)
	S@99R	12.41 (47.9)	13.70 (50.4)	18.44 (57.7)	10.23 (43.1)	**24.54** (86.0)
	R@95S	37.62 (89.3)	37.74 (89.3)	**42.04** (91.1)	38.37 (89.5)	38.90 (90.7)
DR	AUC	87.41 (98.3)	86.41 (98.3)	**88.54** (98.7)	84.45 (98.3)	88.24 (98.7)
	S@95R	37.85 (89.4)	32.30 (87.8)	41.96 (91.1)	31.08 (87.4)	**42.72** (91.2)
	S@99R	8.08 (38.3)	6.31 (33.5)	**14.20** (51.3)	7.77 (38.3)	10.47 (44.6)
	R@95S	54.04 (92.3)	53.69 (92.3)	**57.67** (93.4)	47.88 (91.4)	55.52 (93.1)
PM	AUC	93.37 (98.6)	92.95 (98.6)	93.92 (98.7)	94.09 (98.6)	**94.35** (98.7)
	S@95R	59.62 (93.0)	46.73 (91.2)	60.98 (93.7)	60.61 (93.1)	**67.68** (94.3)
	S@99R	17.59 (56.6)	6.42 (33.9)	28.40 (67.0)	22.88 (62.0)	**45.59** (75.9)
	R@95S	73.07 (94.2)	74.54 (94.3)	**75.92** (94.9)	75.06 (94.3)	74.95 (94.8)
RVO	AUC	89.67 (98.4)	90.37 (98.4)	86.27 (98.3)	85.16 (98.3)	**90.86** (98.7)
	S@95R	20.30 (58.3)	29.92 (67.4)	1.51 (11.5)	69.18 (93.9)	**72.94** (94.7)
	S@99R	13.53 (50.1)	15.63 (53.7)	0.93 (7.8)	27.74 (87.4)	**33.12** (70.3)
	R@95S	70.98 (94.0)	**77.22** (94.5)	70.12 (94.0)	57.87 (92.8)	76.33 (94.9)

marily refines calibration in the high-sensitivity tail. RETFound's RVO Spe@95R jumps dramatically, showing that some retina-SSL representations require full adaptation. Some models show tail-metric degradation after FT, demonstrating that AUC-centric tuning can mask clinically relevant regressions at high-recall operating points, suggesting the need for tail-aware objectives or constrained tuning strategies.

3.2 Real-world Generalization

Geographical Robustness and Tail Performance. Across diseases and models, Performance drop is moderate in AUC but amplified at screening tails. It is demonstrated that ranking ability is relatively stable, but operating-point behavior shifts under geography, demonstrating calibration can move substantially under natural shifts. VLMs with medical/ophthalmic pretraining show strong spatial robustness at screening-relevant points. MedSigLIP consistently achieves the best tail specificity under geographic shift. A plausible driver is that MedSigLIP is trained on de-identified medical image–text pairs, including ophthalmology, improving invariances relevant to clinical image.

Practical Deployment Implications. Our results demonstrate that AUC is a deceptive proxy for geographic readiness. Models with nearly identical AUC values exhibit divergent $S@99R$ profiles, directly translating to disparate referral burdens and specialist workloads at fixed sensitivities. Because much of the degradation concentrates at operating points rather than ranking, lightweight post-hoc recalibration or region-specific threshold selection using a small target-region validation set can yield large gains without retraining the full model, an actionable path for resource-limited deployment.

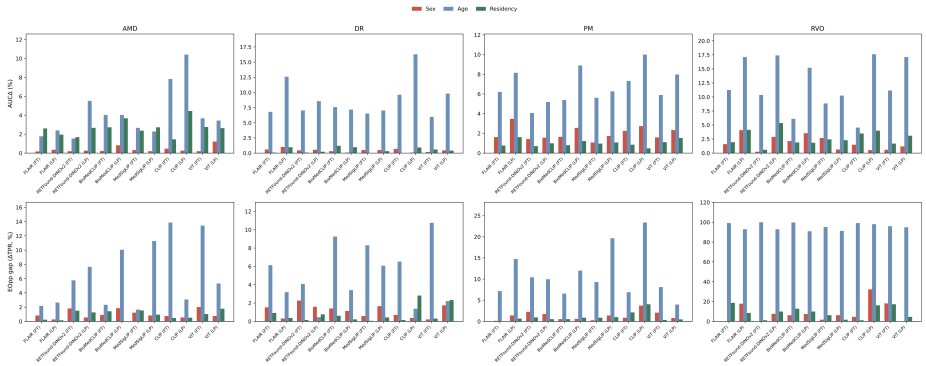

Fig. 2. Fairness disparities across diseases, adaptation strategies, and sensitive attributes.

3.3 Fairness and Socioeconomic Bias

Age-Based Disparities and Metric Incompatibility. Our analysis identifies age as the primary axis of algorithmic bias across all evaluated pathologies and models. Under a unified global operating threshold, age-stratified subgroups exhibit significantly higher EOpp violations compared to sex or insurance status. These findings demonstrate that a fixed operating point systematically redistributes false-positive and false-negative burdens between age strata, reinforcing the necessity of evaluating fairness under clinically intended decision rules rather than relying solely on threshold-independent metrics.

The Fine-Tuning Paradox: Separability vs. Calibration. We observe a critical trade-off between model adaptation and deployment equity: while FT improves overall group-wise separability, it frequently exacerbates parity violations at fixed thresholds. Compared to LP, FT consistently reduces ΔAUC across all model categories, indicating a superior ability to learn disease-specific features for marginalized groups. However, EOpp gaps under a single global threshold increase for FT in multiple categories, suggesting that fine-tuning can alter calibration and score distributions unevenly across groups. This observation is consistent with known trade-offs and incompatibilities among fairness notions when subgroup-based rates differ.

4 Conclusions

We presented **OphFM-bench**, a national-scale benchmark for early retinopathy screening that evaluates foundation models under realistic clinical heterogeneity. We showed that AUC does not necessarily translate to screening efficiency at high-sensitivity operating points, and that geographic shift can disproportionately degrade tail metrics relative to AUC. We further demonstrated that fairness conclusions depend strongly on the deployment decision rule: fine-tuning may improve separability while worsening equalized opportunity under a single global

threshold. Overall, OphFM-bench provides an actionable protocol equalizedfor selecting and adapting FMs for real-world telemedicine screening, emphasizing operating-point validation, cross-region stress testing, and decision-rule-aware bias auditing before deployment.

Acknowledgments. This study was supported by the National Natural Science Foundation of China (No. 82388101, 82471130, 82373681, 82173613), National Clinical Key Specialty Construction Project (No. 10000015Z155080000004).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Chen, J., Zheng, H., Bei, J.X., Sun, L., Jia, W.h., Li, T., Zhang, F., Seielstad, M., Zeng, Y.X., Zhang, X., et al.: Genetic structure of the han chinese population revealed by genome-wide snp variation. The American Journal of Human Genetics **85**(6), 775–785 (2009)
2. Chouffani El Fassi, S., Abdullah, A., Fang, Y., Natarajan, S., Masroor, A.B., Kayali, N., Prakash, S., Henderson, G.E.: Not all ai health tools with regulatory authorization are clinically validated. Nature Medicine **30**(10), 2718–2720 (2024)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
4. Du, J., Guo, J., Zhang, W., Yang, S., Liu, H., Li, H., Wang, N.: Ret-clip: A retinal image foundation model pre-trained with clinical diagnostic reports. In: International conference on medical image computing and computer-assisted intervention. pp. 709–719. Springer (2024)
5. Handler, R., Sharma, S., Hernandez-Boussard, T.: The fragile intelligence of gpt-5 in medicine. Nature Medicine pp. 1–3 (2025)
6. Jeong, Y.D., Park, S., Kim, M.S., Hong, S.H., Aalruz, H., Abate, Y.H., Abbasgholizadeh, R., Abd ElHafeez, S., Abdullahi, A., Aboagye, R.G., et al.: Global burden of vision impairment due to age-related macular degeneration, 1990–2021, with forecasts to 2050: a systematic analysis for the global burden of disease study 2021. The Lancet Global Health **13**(7), e1175–e1190 (2025)
7. Krause, J., Gulshan, V., Rahimy, E., Karth, P., Widner, K., Corrado, G.S., Peng, L., Webster, D.R.: Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. Ophthalmology **125**(8), 1264–1272 (2018)
8. Maulvi, F.A., Desai, D.T., Kalaiselvan, P., Shah, D.O., Willcox, M.D.: Current and emerging strategies for myopia control: a narrative review of optical, pharmacological, behavioural, and adjunctive therapies. Eye **39**(14), 2635–2644 (2025)
9. Moradi, M., Shah, R., Fujita, A., Bineshfar, N., Vu, D.M., Aziz, K., Liebman, D.L., Hashemabad, S.K., Wang, M., Elze, T., et al.: Clinically informed semi-supervised learning improves disease annotation and equity from electronic health records: a glaucoma case study. npj Digital Medicine (2025)

10. Qiu, J., et al.: Development and validation of a multimodal multitask vision foundation model for generalist ophthalmic artificial intelligence. NEJM AI **1**(12), AIoa2300221 (2024). https://doi.org/10.1056/AIoa2300221, https://ai.nejm.org/doi/full/10.1056/AIoa2300221
11. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
12. Rashidisabet, H., Sethi, A., Jindarak, P., Edmonds, J., Chan, R.P., Leiderman, Y.I., Vajaranant, T.S., Yi, D.: Validating the generalizability of ophthalmic artificial intelligence models on real-world clinical data. Translational Vision Science & Technology **12**(11), 8–8 (2023)
13. Rose, S.: Ai hype cycles and reality in health care. In: JAMA Health Forum. vol. 6, pp. e251904–e251904. American Medical Association (2025)
14. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
15. Silva-Rodriguez, J., Chakor, H., Kobbi, R., Dolz, J., Ayed, I.B.: A foundation language-image model of the retina (flair): Encoding expert knowledge in text supervision. Medical Image Analysis **99**, 103357 (2025)
16. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
17. Ting, D.S.W., Cheung, C.Y.L., Lim, G., Tan, G.S.W., Quang, N.D., Gan, A., Hamzah, H., Garcia-Franco, R., San Yeo, I.Y., Lee, S.Y., 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**(22), 2211–2223 (2017)
18. Vujosevic, S., Aldington, S.J., Silva, P., Hernández, C., Scanlon, P., Peto, T., Simó, R.: Screening for diabetic retinopathy: new perspectives and challenges. The Lancet Diabetes & Endocrinology **8**(4), 337–347 (2020)
19. Wang, M., Lin, T., Lin, A., Yu, K., Peng, Y., Wang, L., Chen, C., Zou, K., Liang, H., Chen, M., et al.: Enhancing diagnostic accuracy in rare and common fundus diseases with a knowledge-rich vision-language model. Nature communications **16**(1), 5528 (2025)
20. Wang, T.W., Luo, W.T., Tu, Y.K., Chou, Y.B., Wu, Y.T.: Systematic review and meta-analysis of regulator-approved deep learning systems for fundus diabetic retinopathy detections. npj Digital Medicine (2025)
21. Warraich, H.J., Tazbaz, T., Califf, R.M.: Fda perspective on the regulation of artificial intelligence in health care and biomedicine. Jama **333**(3), 241–247 (2025)
22. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)
23. Zhang, Z., Zhang, H., Pan, Z., Bi, Z., Wan, Y., Song, X., Fan, X.: Evaluating large language models in ophthalmology: systematic review. Journal of Medical Internet Research **27**, e76947 (2025)
24. Zhou, Y., Chia, M.A., Wagner, S.K., Ayhan, M.S., Williamson, D.J., Struyven, R.R., Liu, T., Xu, M., Lozano, M.G., Woodward-Court, P., et al.: A foundation model for generalizable disease detection from retinal images. Nature **622**(7981), 156–163 (2023)