How Generalizable are Foundation Models when applied to Different Demographic Groups and Settings?

Zhuxin Xiong BEng[1], Xiaofei Wang PhD[1*], Yukun Zhou PhD[2,3,4], Pearse A. Keane PhD[3,5], Yih Chung Tham PhD[6,7,8], Ya Xing Wang PhD[9] and Tien Yin Wong PhD[7,10]

1. Key Laboratory for Biomechanics and Mechanobiology of Ministry of Education, Beijing Advanced Innovation Center for Biomedical Engineering, School of Biological Science and Medical Engineering, Beihang University, Beijing, China.
2. Centre for Medical Image Computing, University College London, London, UK.
3. NIHR Biomedical Research Centre at Moorfields Eye Hospital NHS Foundation Trust, London, UK.
4. Department of Medical Physics and Biomedical Engineering, University College London, London, UK.
5. Institute of Ophthalmology, University College London, London, UK.
6. Centre for Innovation and Precision Eye Health, Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore.
7. Singapore Eye Research Institute, Singapore National Eye Centre, Singapore.
8. Ophthalmology and Visual Science Academic Clinical Program, Duke-NUS Medical School, Singapore.
9. Beijing Institute of Ophthalmology, Beijing Ophthalmology and Visual Science Key Lab, Beijing Tongren Eye Center, Beijing Tongren Hospital, Capital Medical University, Beijing, China.
10. Tsinghua Medicine, Tsinghua University, Beijing, China.

Correspondence:

Xiaofei Wang, School of Biological Science and Medical Engineering, Beihang University, Beijing, China. Email: xiaofei.wang@buaa.edu.cn.

Notes:

Dr. Xiaofei Wang, Dr. Yih Chung Tham and Dr. Ya Xing Wang contributed equally to this article.

Abstract

RETFound is a retinal photo-based foundational model that can be fine-tuned to downstream tasks. However, its generalizability to Asian populations remains unclear. In this study, we fine-tuned RETFound on Asian-specific dataset. We then evaluated the performance of RETFound versus conventional Vision Transformer model (pretrained in ImageNet) in diagnosing glaucoma, coronary heart disease, and predicting the 3-year risk of stroke in Asian population. When fine-tuned on 'full' dataset, RETFound showed no significant improvement compared to conventional Vision Transformer model (AUCs of 0.863, 0.628 and 0.557 versus 0.853, 0.621 and 0.543, respectively, all $P \geq 0.2$). Furthermore, in scenarios with very limited training data (fine-tuned on ≤ 25% of the full dataset), RETFound showed a slight advantage (up to a maximum AUC increase of 0.03). However, these improvements were not statistically significant (all $P \geq 0.2$). These findings indicate the challenges foundational AI models face in adapting to diverse demographics, emphasizing the need for expansion of current foundation models to include data of greater diversity and highlighting the necessity of global collaboration on foundation model research.

The recent emergence of foundational artificial intelligence (AI) models offers a promising solution for addressing the significant costs of collecting and annotating large datasets in training clinically usable medical AI models for disease detection.[1] In theory, foundation models are neural networks trained on unlabeled datasets that can be adapted to handle a wide variety of downstream tasks and in different settings without the need to re-train the model from scratch. RETFound developed by Zhou et al.[1] was such a foundation model that has set a benchmark for broad-spectrum ocular and systemic disease detection using retinal fundus photos and optical coherence tomography images. In this groundbreaking work, RETFound showed excellent performance in predicting ocular and systemic diseases by finetuning even with limited labelled data. As a foundational model, RETFound was trained on a vast number of retinal images, making it highly specialized for tasks involving retinal features. However, it remains unclear if foundational AI models such as RETFound are truly generalizable to other demographic groups and different settings. The aim of our study was to test the generalizability of the RETFound across a different demographic group, through three distinct tasks in Asian populations: diagnosis of glaucoma, diagnosis of coronary heart disease, and prediction of 3-year risk of incident stroke.

To assess the generalizability of the RETFound, we conducted three testing tasks similar to those outlined in the RETFound paper but with a different population. In the original paper, the foundation model RETFound was developed first with a large amount of unlabeled retinal images data mainly from Moorfields Diabetic imAge dataSet (MEH-MIDAS), then underwent fine-tuning for specific tasks such as glaucoma detection. Similarly, we fine-tuned the pretrained RETFound using our Asian population dataset with different characteristics (**Table S1 in the Supplementary Appendix**) and performed three tasks matching the tasks in the original paper: glaucoma diagnosis,

cardiovascular disease diagnosis, and three-year stroke prediction. Additionally, to eliminate any bias related to model architecture, we used the Vision Transformer[2] model (ViT-large, the same model architecture as RETFound's encoder) pretrained by means of supervised learning on ImageNet-1k (about 1.4 million natural images) as a reference to examine the performance improvement achieved by the RETFound across varying data volumes for each task (**Figure1**). In each task, the models were adapted with labelled training data, and evaluated on held-out internal test sets. Both the glaucoma and stroke prediction tasks were identical to those in the RETFound paper. For cardiovascular disease prediction, we performed a cross-sectional prediction of coronary heart disease, while the RETFound paper focused on longitudinal predictions of 3-year heart failure and myocardial infarction. In the original paper, RETFound exhibited excellent efficacy across these three tasks.

The data for the initial two tasks were obtained from the a health examination cohort of 161,943 participants, which is an ongoing cohort study of healthy individuals who undergo biannual examinations.[3] For prediction of 3-year incident stroke, data were obtained from the another follow-up study[4] in which participants were examined annually for the occurrence of stroke over a continuous five-year period. Details on our datasets and evaluation methods can be found in supplemental materials (**Supplementary Methods in the Supplementary Appendix**). Model performance was reported using their respective best Area Under the ROC Curve (AUC) values, recall values, precision values and F1 scores. We employed five-fold cross-validation for each experiment, yielding five AUC values, recall values, precision values and F1 scores per task. The means and the 95% confidence interval (CI) were reported. P-values were calculated using paired two-sided t-tests between the RETFound and the Vision Transformer model for each task, similar to the methods used in the RETFound paper.

We found no significant improvement when using RETFound on our datasets compared to the Vision Transformer model pretrained on natural images in all three tasks (**Figure2** and **Table1**). For detecting glaucoma, AUCs of 0.863 (95% CI - 0.831 to 0.895), 0.846 (95% CI - 0.818 to 0.873), 0.851 (95% CI - 0.812 to 0.890) and 0.838 (95% CI - 0.807 to 0.869) were achieved by fine-tuning the RETFound model on the entire dataset, 50%, 25% and 10% of the dataset, respectively. In contrast, the Vision Transformer model yielded AUC values of 0.853 (95% CI - 0.826 to 0.880), 0.836 (95% CI - 0.789 to 0.882), 0.821 (95% CI - 0.788 to 0.854) and 0.808 (95% CI - 0.718 to 0.899). The RETFound model showed a slight improvement in all the AUCs (with the highest AUC improvement being 0.03), especially with limited training data, but the results were not statistically different (all $P \geq .3$).

In the context of coronary heart disease detection, fine-tuning the RETFound resulted in AUC values of 0.628 (95% CI - 0.621 to 0.634), 0.622 (95% CI - 0.607 to 0.637) and 0.600 (95% CI - 0.583 to 0.617) for the entire dataset, 50% and 10% of the dataset, respectively, exhibiting negligible discrepancies relative to the AUC values of 0.621 (95% CI - 0.613 to 0.628), 0.610 (95% CI - 0.582 to 0.638) and 0.579 (95% CI - 0.567 to 0.591) obtained using the Vision Transformer model (all $P \geq .2$). For 3-year stroke prediction, the best AUC value achieved by fine-tuning the RETFound was 0.557 (95% CI - 0.533 to 0.581), compared to 0.543 (95% CI - 0.497 to 0.590) obtained using the Vision Transformer ($P = .712$). In terms of additional evaluation metrics such as recall values, precision values, and F1 scores, the RETFound model's performance on all datasets for the three tasks demonstrated no notable differences over the Vision Transformer model (all $P \geq .04$). All quantitative results are listed in supplemental materials (**Table S2 in the Supplementary Appendix**).

Our study tested the generalizability and usability of RETFound, a new foundational AI model trained on mostly UK data, on Asian datasets. We found

no significant performance gains using RETFound compared to the Vision Transformer model across three ocular and systemic disease detection tasks. Notably, the best AUC increase observed in our study (0.014) was considerably lower than the improvement reported in the original RETFound paper (0.126) for similar tasks. While RETFound showed marginally better performance in scenarios with limited training data (25% of the whole training data), the improvement remained negligible in our Asian cohorts (best AUC increase: 0.03). This consistent underperformance across multiple tasks raises concerns about how foundational AI models can be easily applied to other demographic groups and healthcare settings without the need for substantial training using local data.

Subgroup analyses based on gender and age were conducted to evaluate the performance of the RETFound model (**Supplementary Methods, Table S3 and Table S4**). The first analysis assessed classification performance across gender and age groups, revealing no significant differences between genders but better performance in older age groups. The second analysis compared RETFound and Vision Transformer within these subgroups, showing no statistically significant improvement for RETFound compared to Vision Transformer in any subgroup. Specifically, while RETFound performed better in older age groups, the coronary heart disease and stroke tasks, which had age profiles similar to the pretraining dataset (average age 64.5 ± 13.3 years), did not show significant performance improvement over Vision Transformer. Similarly, the glaucoma dataset, with a younger age profile, also showed no significant improvement when divided into younger and older groups These findings suggest that factors beyond age and gender representation, such as ethnicity, may contribute to the lack of performance improvement observed.

Our paper underscores the continuing need to have diverse demographic representation in training data even for foundational AI models. This limitation highlights a broader but critical issue in AI development for healthcare

applications. The inclusion of data from multiple ethnicities and settings has always been essential to ensure the efficacy of current AI models across diverse demographic groups.[5] Our study now demonstrates that even so-called foundational AI models are not immune to the need for training on diverse demographic groups and testing in different real-world settings.

In conclusion, while foundation AI models promises to be a significant advancement in medical AI, their current limitations remain similar to those of traditional AI models. The generalizability and applicability of foundational AI models to different demographic groups and clinical context are limited unless foundational models are also trained on more diverse datasets and tested in various clinical settings. Our study highlights the necessity of global collaboration on foundation model research to significantly incorporate a more diverse dataset during the training phase of foundational AI models. This step is vital for developing foundational AI models that are not only efficacious but effective, equitable and generalizable, enabling them to tackle various diseases and clinical problems in global healthcare settings.

References

1. Zhou Y, Chia MA, Wagner SK, et al. A foundation model for generalizable disease detection from retinal images. *Nature*. 2023;622(7981):156-163. DOI:10.1038/s41586-023-06555-x.
2. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Published online 2020. DOI:10.48550/ARXIV.2010.11929.
3. Xue CC, Gao LQ, Cui J, et al. Gunn's dots as indicators of renal function, findings from the Tongren Health Care Study. *Retina*. 2022;42(4):789-796. DOI:10.1097/IAE.0000000000003354.
4. Tian X, Zuo Y, Chen S, et al. Triglyceride–glucose index is associated with the risk of myocardial infarction: an 11-year prospective study in the Kailuan cohort. *Cardiovasc Diabetol*. 2021;20(1):19. DOI:10.1186/s12933-020-01210-5.
5. Zou J, Schiebinger L. Ensuring that biomedical AI benefits diverse populations. *EBioMedicine*. 2021;67:103358. DOI:10.1016/j.ebiom.2021.103358.

Table1. Summary of experimental datasets and results

Tasks	***Positive + Negative of Datasets, No. (%)***	***RETFound AUC, mean (95% CI)***	***Vision Transformer AUC, mean (95% CI)***	***AUC improvements, mean***	***P value***
Glaucoma	1512 + 1576 (100)	0.863 (0.831, 0.895)	0.853 (0.826, 0.880)	0.010	.450
	756 + 788 (50)	0.846 (0.818, 0.873)	0.836 (0.789, 0.882)	0.011	.690
	378 + 394 (25)	0.851 (0.812, 0.890)	0.821 (0.788, 0.854)	0.030	.319
	151 + 158 (10)	0.838 (0.807, 0.869)	0.808 (0.718, 0.899)	0.030	.557
Coronary Heart Disease	1775 + 3550 (100)	0.628 (0.621, 0.634)	0.621 (0.613, 0.628)	0.007	.276
	888 + 1776 (50)	0.622 (0.607, 0.637)	0.610 (0.582, 0.638)	0.012	.345
	355 + 710 (20)	0.600 (0.583, 0.617)	0.579 (0.567, 0.591)	0.021	.219
Stroke	261 + 1305 (100)	0.557 (0.533, 0.581)	0.543 (0.497, 0.590)	0.014	.712

Figure Legends

Figure1. Overview of the study

Figure2. Performance of RETFound versus conventional Vision Transformer in 3 tasks.Models are adapted to each dataset in different data volume by fine-tuning and internally evaluated on hold-out test data in the tasks of diagnosing glaucoma, coronary heart disease and predicting the three-year risk of stroke. The error bars show 95% CI and the bar centre represents the mean value of the AUC of the five-fold cross-validation. We compare the performance of RETFound with the conventional Vision Transformer to check whether statistically significant differences exist. P value is calculated with the paired two-sided t-test and listed in the figure.