# A new foundation model for multimodal ophthalmic images: advancing disease detection and prediction

| | |
|---|---|
| Journal: | *NEJM AI* |
| Manuscript ID | Draft |
| Manuscript Type: | Perspective |
| Date Submitted by the Author: | n/a |
| Complete List of Authors: | Chia, Mark ; NIHR Moorfields Biomedical Research Centre<br>Zhou, Yukun; UCL, Centre for Medical Image Computing<br>Keane, Pearse; University College London Institute of Ophthalmology, |
| | |

# A new foundation model for multimodal ophthalmic images: advancing disease detection and prediction

## Mark A Chia, Yukun Zhou, Pearse A Keane

1. University College London, 11-43 Bath Street, London EC1V 9EL, United Kingdom
2. Moorfields Eye Hospital, 162 City Road, London EC1V 2PD, United Kingdom

## Correspondence

Pearse Keane, p.keane@ucl.ac.uk, 11-43 Bath Street, London EC1V 9EL, United Kingdom

## Conflicts of Interest

MAC and YZ report no financial disclosures. PAK has acted as a consultant for DeepMind, Roche, Novartis, Apellis, and BitFount and is an equity owner in Big Picture Medical. He has received speaker fees from Heidelberg Engineering, Topcon, Allergan, and Bayer

## Paper Link

AI-23-00221.R6-Proof-hi.pdf

## Requirements

- 1-2 sentence abstract
- Less than 1000 words
- No figures/tables/references
- Accessible and of interest to non-specialists
- Due October 13th

# Abstract

Foundation models are a powerful tool in ophthalmology for building generalisable systems that can be efficiently applied to a range of ocular and systemic health tasks. A new foundation model for ophthalmic images demonstrates important progress, particularly through its flexible approach to multimodal training and its application to image segmentation tasks.

Over the last decade, rapid developments in deep learning have sparked growing enthusiasm for artificial intelligence (AI). Ophthalmology has been at the forefront of these advances within medicine, with the approval of the first two autonomous AI systems by the Food and Drug Administration coming from the specialty [1]. More recently, AI has seen a further surge in excitement, primarily due to the advent of large-scale foundation models. AI's remarkable abilities have gained mainstream recognition with the release of generative foundation models such as ChatGPT and Stable Diffusion.

The term 'foundation model' was introduced by researchers at Stanford's Human-Centred AI Institute in 2021 [2]. It refers to large AI models trained on extensive datasets, which can then be fine-tuned for numerous specific tasks. Although foundation models share many commonalities with deep learning and transfer learning methods, they represent a significant departure from traditional AI in terms of their sheer scale and broad potential uses. Earlier AI models were typically built to handle one specific task, whereas foundation models are designed as flexible tools that can be applied to a range of problems. Within medicine, their development has been enabled by advances in two key areas. First, a training approach known as self-supervised learning has allowed researchers to tap into the vast quantities of unlabelled data which accumulate during routine medical practice. Second, novel model architectures known as vision transformers, alongside exponential growth in computational power, have allowed scaling of models beyond what was previously possible. Released in 2023, RETFound became the first foundation model in ophthalmology, validated on 13 tasks spanning retinal disease diagnosis, retinal disease prediction, and the prediction of future systemic events [3].

In NEJM AI, Qiu et al. present novel work on the development and validation of VisionFM, a new foundation model for ophthalmic images *[insert Qiu et al reference]*. VisionFM was trained on 3.4 million ophthalmic images from over half a million patients, sourced from a combination of public and private datasets. The training approach utilised a vision transformer architecture and employed a self-supervised learning method known as iBOT, which learns representations by filling in missing parts of images. The authors subsequently validated VisionFM on various tasks, including ophthalmic disease classification, prediction of glaucoma progression, identification of systemic diseases, and image segmentation.

For retinal disease classification, the authors explored four key areas and reported impressive results that demonstrate the generalizability of VisionFM. These areas included: (1) internal performance compared to competitor models for classifying eight common retinal conditions using different imaging modalities, (2) internal performance compared to junior and intermediate-level ophthalmologists in classifying eye diseases from retinal photographs, (3) external performance on a subset of tasks, and (4) external performance using a new imaging modality (optical coherence tomography angiography) and a new disease (ocular albinism) that were not encountered during pre-training.

Qiu et al. also report capable performance on more complex tasks, such as predicting glaucoma progression and identifying systemic health issues using fundus photos. The two tasks explored for systemic health linkage were predicting the results of a panel of blood biomarkers and identifying the presence of an intracranial tumour. Perhaps the most noteworthy development of VisionFM was its performance on a range of image segmentation tasks, including vessel segmentation on fundus photos and layer segmentation on optical coherence tomography.

The progress in foundation models achieved through this work can be summarised in two key areas. First, Qiu et al. demonstrate a pre-training approach that allows for the flexible integration of multiple imaging modalities into a single model, with modality-agnostic fine-tuning capabilities. Second, the authors show that the versatile performance of foundation models can be extended to a variety of useful image segmentation tasks, which have been previously underexplored.

Despite this progress, it is important to acknowledge several relevant limitations. First, only the simpler retinal disease classification tasks were validated on external datasets, while the remaining tasks were evaluated on internal sets, which have known drawbacks related to domain shift and overfitting. Second, the validation tasks often involved curated datasets that required unanimous agreement for inclusion, likely excluding complex cases and thus exaggerating performance compared to a real-world spectrum of disease. Third, the retinal disease classification task assumed a single, prominent disease within each image; however, it is probable that the images contained multiple diseases due to known associations between conditions. Fourth, when comparing VisionFM to clinicians, specialists were often evaluated under artificial conditions, such as being asked to assess a single image without access to relevant patient history and exam findings, which are typically available. Fifth, the evaluation of generalisability to an unseen eye disease was limited to classifying ocular albinism, which typically exhibits an obvious clinical phenotype and may not reflect more subtle conditions. Lastly, while VisionFM's performance in identifying systemic diseases was impressive, considerable work remains to demonstrate its real-world clinical utility in these cases.

To fully realise the potential of foundation models in ophthalmology, it is important to explore several key avenues. Our current understanding of how the characteristics of training data influence the downstream performance of these models remains limited. One of the most promising aspects of foundation models is their potential to address equity concerns by facilitating robust performance, even with limited data, particularly for rare eye conditions or patients from underrepresented ethnic groups. Gaining deeper insights into the characteristics of training data is a crucial step toward this goal. Additionally, while multimodal image integration is vital, the true benefits of foundation models are likely to be realised only through the seamless incorporation of multiple data types, including text, audio, and genomic information. Much like a clinician, a foundation model should integrate these diverse data types using a unified strategy. Finally, true generalizability will be enhanced by employing globally diverse datasets. Achieving this requires not only strong international collaborations but also innovative technical solutions, including the exploration of privacy-preserving strategies such as federated learning. With further investigation, foundation models have the potential to transform healthcare, ultimately leading to more personalised and equitable patient care.

# References

1. Chia MA, Antaki F, Zhou Y, Turner AW. Foundation models in ophthalmology. Available: https://bjo.bmj.com/content/early/2024/06/04/bjo-2024-325459.abstract

2. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. On the Opportunities and Risks of Foundation Models. arXiv [cs.LG]. 2021. Available: http://arxiv.org/abs/2108.07258

3. Zhou Y, Chia MA, Wagner SK, Ayhan MS, Williamson DJ, Struyven RR, et al. A foundation model for generalizable disease detection from retinal images. Nature. 2023;622: 156–163.