General purpose AI: Transitioning from high performance in highly curated settings to useful applications in ophthalmology

BY ARUN JAMES THIRUNAVUKARASU

Artificial intelligence (AI) may be understood as the ability of machines to perform tasks which otherwise require human perception, reasoning, or learning. With the advent of deep learning, AI has achieved remarkable results across a wide range of medical tasks [1]. However, skilled clinicians and conventional infrastructure remain the backbones of healthcare and most ophthalmology patients outside research studies still do not benefit from AI tools [2]. Most proof-of-concept studies are not resulting in clinical deployment due to limited utility outside highly curated settings. However, emerging technology is resulting in models with generalisable abilities rather than being trained in a single, narrow task. There remain significant barriers to implementation, but these can be reframed as areas with the greatest potential for positive change. By addressing ethical issues, improving practical systems design, and enhancing research, AI can fulfill its potential to revolutionise eyecare.

Conventional models exhibit high performance but low versatility

Various ophthalmological AI models have received Food & Drug Administration (FDA) approval, and the first autonomous model across medicine to be awarded FDA approval was a diabetic retinopathy (DR) screening application [3,4]. This progress is remarkable, but there remains significant disparity between the thousands of models reported in the literature and those available for patient care, due to limited versatility of AI models in two domains: algorithmic and population based.

Algorithmic versatility is limited by 'supervised' training schemata employed by AI developers. Supervised learning depends on the narrow definition of a task such as classifying normal and disease-associated images, and training using images labelled by qualified clinicians [5]. Using this accurate and reliable 'ground truth', AI models learn how to differentiate between classes independently and thereby attain comparable performance to clinicians. Many models with strong diagnostic performance in age-related macular degeneration, DR, glaucoma, and other conditions have been developed in this manner. Moreover, exploratory investigation using classes for which ophthalmological features are unknown has resulted in novel AI applications for prognostication, extraocular risk stratification, and identification of characteristics such as biological sex [6–8]. However, these models are not easily combined into a single application, meaning that use cases are as narrow as the initially defined task: a severe limitation relative to qualified ophthalmologists.

Supervised learning is also limited in terms of population-based versatility by the scope and quality of the data used for training. Performance may not generalise to broader populations if certain subgroups are underrepresented in the training data, and this has been shown to lead to biased output and unequal accuracy [9]. At a global level, high-quality ophthalmological data have only been sourced from a few high-income countries, and disproportionate representation of demographics within datasets compounds existing disparities [9–11]. Overcoming these algorithmic and population-based limitations is a significant challenge requiring large quantities of accurately labelled data from representative samples. The largest available datasets relate to fundus photography, particularly for DR [11]. It is therefore unsurprising that most FDA-approved models relate to fundus photography, and to DR [3].

The potential of foundation models

Mitigating the limitations of supervised learning is one way to improve the clinical utility of AI in ophthalmology. Multimodal foundation models—AI applications with general ability to interpret, process, and produce data in many formats—are emerging, beginning with large language models (LLMs) such as GPT-4, Gemini, and LLaMA 3; as well as image-based medical foundation models such as RETFound [12,13]. Large language models are pre-trained on large quantities of text to learn how language is formulated before being fine-tuned through grading of model outputs in response to queries. Image-based foundation models are developed in a similar way, with pretraining using enormous numbers of images followed by fine-tuning on more specific tasks. Large language models have attained impressive medical examination results and early experiments suggest that LLMs' responses to a wide range of queries compare well with expert ophthalmologists [14]. The next step is combining text-based reasoning with image interpretation within 'vision-language models'.

Foundation models could soon be able to contribute significantly to many aspects of clinical work with appropriate oversight: automating triage, assisting with diagnosis and management, as well as augmenting medical education by providing on-demand interactive expertise. Autonomous deployment of LLMs is currently precluded by their propensity to fabricate facts ('hallucinate') and inability to accurately express uncertainty [15]. Deployment in carefully defined roles with clinician supervision may be feasible, but such applications must fulfil the same requirements as conventional models to be implemented: robust validation, bias mitigation, and post-deployment governance [12,16]. Despite these barriers, great change in ophthalmology seems feasible. Large language models may leverage guidelines and textbooks to direct decision-making, and foundation models may be fine-tuned on specific populations to minimise disadvantage caused by algorithmic bias. Patients could converse with a chatbot before being triaged to eye units based on the urgency of their presentation; have vision tests and imaging interpreted by AI to make a diagnosis and suggest management decisions; and be discharged with initial advice and ability to have questions answered by chatbot on-demand. Supervising ophthalmologists may see fewer patients or require less time per appointment, helping to address severely prolonged waiting lists and improve the quality of care by allowing ophthalmologists to focus on communication and clinical skills rather than administrative tasks and unnecessary appointments [17].

Barriers to implementation are opportunities for development

Significant barriers are preventing implementation of available technology and these also affect emerging foundation models. However, these barriers represent the areas with the most potential for innovation: where development could lead to impactful changes in ophthalmology and general healthcare.

First, ethical concerns must be addressed. Fundamentally, stakeholders must agree on what roles are appropriate for AI to take on. Should models be allowed to make decisions about care without clinician-oversight? What standard of validation is required to justify deployment of an AI decision-aid? Do patients have a right to care without AI intervention? As advanced AI applications remain relatively nascent in ophthalmology, there is time to answer these questions and design the future, but patients and practitioners must be consulted openly for a consensus to be reached. Additionally, bias and fairness concerns represent significant risks to affected populations, such as suboptimal performance in subgroups that are underrepresented in the data used for training and validation [9]. These require purposeful engineering and greater transparency with models intended for use in patient care [18,19]. Encouraging use of open-source models such as LLaMA 3 and RETFound will help maximise transparency and may also help address bias by facilitating specific local training or fine-tuning with data from relevant populations.

The formidable infrastructural requirement to deploy the most advanced AI models is another challenge. In addition to being prohibitively expensive for most healthcare institutions and even most countries, these enormous computational arrays are associated with a significant and growing carbon footprint [20]. As societies move towards reducing energy demand to mitigate climate change, difficult decisions must be made about which projects to support. To avoid perpetuating global inequality it is essential to investigate less energy-intensive and resource-demanding methods of implementing available technology to improve provision of eyecare. Forecasts suggest that the energy costs of AI development will eventually plateau and then fall, but urgent consideration is required while demand continues to grow [21].

Finally, methodological shortcomings must be overcome. Very few AI interventions are pitted against clinical practice in randomised control trials (RCTs); just one ophthalmological study (evaluating an AI model for diagnosis of childhood cataract) was identified in a 2022 systematic review [22,23]. This makes it difficult to establish how new applications compare to conventional practice. Researchers should explicitly describe the purported benefits of an AI intervention and use these as endpoints in pragmatic RCTs to conclusively demonstrate effectiveness [24]. Without this evidence, it is difficult to justify expensive and risky delegation of clinical work to AI.

Does the world want AI ophthalmology?

Technology has progressed to a point where it is conceivable that AI systems may triage, diagnose, and manage ophthalmology patients with supervising human clinicians empowered to deliver a higher standard of care. Ophthalmology has historically been in the vanguard of medical AI development, and this trend seems likely to continue given the rich data collected in usual clinical practice. However, practical and technical limitations mean that radical changes are not yet feasible. While work to improve these applications continues, patients and practitioners must decide how AI applications should and should not contribute to clinical ophthalmology, providing essential direction for how research and development should be directed, governance structures can be established, and clinicians can be educated.

References

1. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;**521**:436–44.
2. Kapoor R, Walters SP, Al-Aswad LA. The current state of artificial intelligence in ophthalmology. *Surv Ophthalmol* 2019;**64**:233–40.
3. U.S. Food and Drug Administration. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices (2022). *FDA.* https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices
4. Abràmoff MD, Lavin PT, Birch M, et al. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med* 2018;**1**:39.
5. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nat Med* 2019;**25**:24–9.
6. Poplin R, Varadarajan AV, Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng* 2018;**2**:158–64.
7. Korot E, Pontikos N, Liu X, et al. Predicting sex from retinal fundus photographs using automated deep learning. *Sci Rep* 2021;**11**:10286.
8. Bridge J, Harding S, Zheng Y. Development and validation of a novel prognostic model for predicting AMD progression using longitudinal fundus images. *BMJ Open Ophthalmol* 2020;**5**:e000569.
9. Larrazabal AJ, Nieto N, Peterson V, et al. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc Natl Acad Sci USA* 2020;**117(23)**:12592–4.
10. Nakayama LF, Kras A, Ribeiro LZ, et al. Global disparity bias in ophthalmology artificial intelligence applications. *BMJ Health Care Inform* 2022;**29**:e100470.
11. Khan SM, Liu X, Nath S, et al. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *Lancet Digit Health* 2021;**3**:e51–66.
12. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. *Nat Med* 2023;**29**:1930–40.
13. Zhou Y, Chia MA, Wagner SK, et al. A foundation model for generalizable disease detection from retinal images. *Nature* 2023;**7981**:156–63.
14. Thirunavukarasu AJ, Mahmood S, Malem A, et al. Large language models approach expert-level clinical knowledge and reasoning in ophthalmology: A head-to-head cross-sectional study. *PLOS Digit Health* 2024;**3(4)**:e0000341.
15. Thirunavukarasu AJ. Large language models will not replace healthcare professionals: curbing popular fears and hype. *J R Soc Med* 2023;**116(5)**:181–2.
16. Gilbert S, Harvey H, Melvin T, et al. Large language model AI chatbots require approval as medical devices. *Nat Med* 2023;**29(10)**:2396–8.
17. Hall R. Hundreds left with lost or damaged eyesight after NHS delays – research (2023). *The Guardian.* https://www.theguardian.com/society/2023/mar/21/hundreds-left-with-lost-or-damaged-eyesight-after-nhs-delays-research
18. Ali MR, Lawson CA, Wood AM, Khunti K. Addressing ethnic and global health inequalities in the era of artificial intelligence healthcare models: a call for responsible implementation. *J R Soc Med* 2023;**116(8)**:260–2.
19. Thirunavukarasu AJ, Elangovan K, Gutierrez L, et al. Democratizing Artificial Intelligence Imaging Analysis With Automated Machine Learning: Tutorial. *J Med Internet Res* 2023;**25**:e49949.
20. Strubell E, Ganesh A, McCallum A. Energy and Policy Considerations for Deep Learning in NLP (2019). http://arxiv.org/abs/1906.02243 [ePub ahead of print]
21. Patterson D, Gonzalez J, Hölzle U, et al. The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink (2022). https://doi.org/10.48550/arXiv.2204.05149 [ePub ahead of print]
22. Lin H, Li R, Liu Z, et al. Diagnostic Efficacy and Therapeutic Decision-making Capacity of an Artificial Intelligence Platform for Childhood Cataracts in Eye Clinics: A Multicentre Randomized Controlled Trial. *eClinicalMedicine* 2019;**9**:52–9.
23. Plana D, Shung DL, Grimshaw AA, et al. Randomized Clinical Trials of Machine Learning Interventions in Health Care: A Systematic Review. *JAMA Netw Open* 2022;**5(9)**:e2233946.
24. Prasad VK, Cifu AS. *Ending Medical Reversal: Improving Outcomes, Saving Lives.* Maryland, USA; Johns Hopkins University Press; 2015.

[All links last accessed May 2024]

SECTION EDITORS

Nima Ghadiri,

Medical Ophthalmology Consultant and Honorary Senior Clinical Lecturer, Liverpool, UK.

Arun James Thirunavukarasu, MA, MB, BChir (Cantab),

Academic Foundation Doctor, Oxford University Hospitals NHS Foundation Trust; Clinical Research Fellow, Nuffield Department of Clinical Neurosciences & Big Data Institute, University of Oxford; Rising Leader Fellow, Aspen Institute, UK.

Declaration of competing interests: Arun James Thirunavukarasu has received research funding from HealthSense for work related to machine learning applications in evidence-based medicine.