The ROYAL COLLEGE of OPHTHALMOLOGISTS

EDITORIAL

Cochrane corner: artificial intelligence for diagnosing exudative age-related macular degeneration

Eye (2025) 39:620–621; https://doi.org/10.1038/s41433-025-03599-3

Exudative or neovascular age-related macular degeneration (nAMD) is one of the leading causes of severe vision loss in older adults worldwide [1] and affects an estimated 2% of Europeans aged over 65 [2]. The potential impact on individuals' and caregivers' quality of life is profound [3]. In addition, nAMD contributes to a significant burden on healthcare resources due to the need for ongoing monitoring and treatment with costly intravitreal anti-vascular endothelial growth factor (anti-VEGF) injections [4], with late diagnosis and delayed treatment initiation producing diminishing returns in terms of long-term visual outcomes [5]. The incidence of nAMD is projected to continue rising as the population ages, making early detection even more essential.

Advancements in artificial intelligence (AI) for retinal image analysis have the potential to improve patient outcomes by enabling early detection and more accurate diagnosis, and hence more timely intervention. This commentary features a recent Cochrane review which evaluates the accuracy of artificial intelligence (AI) tools in diagnosing nAMD [6].

The authors identified 36 eligible diagnostic test accuracy (DTA) studies published up to April 2024 which evaluated 40 AI algorithms. This comprised 16,655 participants across 20 studies analysing optical coherence tomography (OCT) scans, fundus images, infrared images, OCT angiography, or a mix of these. The total cohort size could not be determined as the remaining 16 studies did not report on the number of participants. Demographics were poorly reported as well – only four studies described participants' age and sex, and none reported ethnicity. However, the populations studied did encompass countries across Asia, Europe, and the United States.

Twenty-eight algorithms were internally validated, demonstrating high accuracy with a summary sensitivity of 0.93 (95% confidence interval (CI) 0.89–0.96) and specificity of 0.96 (95% CI 0.94–0.98). A further three underwent external validation, demonstrating similarly excellent performance with a pooled sensitivity of 0.94 (95% CI 0.90–0.97) and a high specificity with wider CIs (0.99, 95% CI 0.76–1.00). The remaining nine studies did not provide data suitable for meta-analysis.

While these AI algorithms appear to perform promisingly well for diagnosing nAMD, these results should be interpreted with caution because of the risk of overfitting with small datasets and internal validation studies, and the low certainty of evidence from imprecision (wide CIs) and risk of bias.

None of the studies were free of bias across the four domains of the modified Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool. For example, multiple studies did not report the numbers and experience levels of human graders setting the reference standard (16/36, 44%), describe masking or independence (35/36, 97%), or whether the graders were provided with clinical information (25/36, 69%). Given the subjective nature of image-based reference standards and the impact of errors on algorithmic performance [7], robust evaluations should involve at least two experienced graders and a robust arbitration process [8]. In addition, it is important to consider whether the reference standard should be based on expert(s) grading the same image as the AI model, or be benchmarked against the clinical gold standard with fluorescein angiography and/or multimodal imaging. For several studies, the image used was sometimes a single modality such a fundus photograph, which would not be used alone in clinical practice for nAMD detection [9].

Study design was another area of concern. The prevalence of nAMD across the studies was artificially high (33%, range 0.3–49%), as the majority (31/36, 86%) employed a case-control design, most of which compared patients with and without nAMD, rather than nAMD versus a spectrum of other retinal diseases with potentially similar imaging features. The former presents a distinct and less complex task than typically seen in clinical practice. In addition, the strict eligibility criteria and exclusion of patients with additional ocular conditions or diagnostic uncertainties may produce unrealistically optimistic results, as such "clean" datasets do not reflect real-world diagnostic challenges. This is especially true for the use case where these models could deliver the highest value proposition - nAMD detection for non-specialists managing populations with a wide variety of complaints. This misalignment between datasets and the potential implementation niche invites spectrum bias and risks inflating AI performance.

Overall, this review highlights a clear need for improved reporting of diagnostic accuracy studies. Reporting standards which cover diagnostic accuracy studies (Standards for Reporting of Diagnostic Accuracy Studies, STARD) and AI studies (Minimal Information about Clinical Artificial Intelligence Modelling, MI-CLAIM) are already well established [10, 11]. The forthcoming STARD-AI extension [12] may help to improve reporting, but will require active implementation from journal editors and other stakeholders given the limited compliance with existing tools.

This Cochrane review has also surfaced inadequate reporting of sociodemographic characteristics across multiple studies, which limits our understanding of the model's performance across diverse patient populations. The MINimum Information for Medical AI Reporting (MINIMAR) standards recommend reporting demographic variables including age, sex, race, ethnicity, and socioeconomic status at a minimum [13]. More recently, the STANDING Together collaboration has developed international consensus recommendations to highlight and/or mitigate bias in datasets used to develop and validate AI models [14]. In addition to reporting relevant patient metadata, the importance of evaluating AI performance across these patient subgroups is emphasized – beyond simply aggregating performance, assessing whether the AI is 'safe on average, or safe for all' is essential [15].

While the lack of external validation studies is concerning, real-world applicability should extend beyond simple dichotomous

Received: 1 December 2024 Revised: 20 December 2024 Accepted: 8 January 2025
Published online: 20 January 2025

concepts of internal versus external validation. Future studies should consider real world deployment such as silent trials (also known as translational trials) [16], randomised controlled trials, or prospective deployment studies with adequate safety guardrails to evaluate algorithm performance in clinical environments. In addition to diagnostic accuracy, this should incorporate evaluations of human-computer interactions and patient-centred outcomes to obtain insights into the system-wide impact of AI models on healthcare services and help build robust evidence for clinical utility and feasibility.

There are several key considerations outside of this Cochrane review. Should an AI model for diagnosing nAMD function autonomously, or serve as a decision support tool for clinicians? Should it be used to triage symptomatic patients in primary care and remote settings where access to specialist care is more limited? What value can these models offer in well-resourced secondary care settings? Such considerations have important implications for evidence generation to support regulatory approval processes, and for other stakeholders such as payers and policymakers as they consider reimbursement structures that facilitate the sustainable provision of patient benefit by AI developers.

This Cochrane review highlights the potential of AI to transform current paradigms of nAMD detection. It also highlights significant gaps in the current evidence base, including inadequate reporting, external validation, and real-world evaluations. Addressing these gaps will require robust study designs, adherence to reporting standards, and greater clarity on how diagnostic AI can fit into the clinical workflow. These are essential steps towards bridging the “AI chasm” [17], and develop early signals of efficacy into products that can be integrated in routine clinical practice to achieve scalable benefit to patients and healthcare services.

Ariel Yuhan Ong [1,2,3,4] ✉, Henry David Jeffry Hogg [1,5,6] and Pearse A. Keane [1,2,4]
[1] Moorfields Eye Hospital NHS Foundation Trust, London, United Kingdom. [2] Institute of Ophthalmology, University College London, London, United Kingdom. [3] Oxford Eye Hospital, Oxford University Hospitals NHS Foundation Trust, Oxford, United Kingdom. [4] NIHR Moorfields Biomedical Research Centre, London, United Kingdom. [5] University Hospitals Birmingham NHS Foundation Trust, Birmingham, United Kingdom. [6] Department of Applied Health Research, School of Health Sciences, College of Medicine and Health, University of Birmingham, Birmingham, United Kingdom.
✉*email: ariel.ong@nhs.net*

REFERENCES

1. Wong WL, Su X, Li X, Cheung CM, Klein R, Cheng CY, et al. Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. Lancet Glob Health. 2014;2:e106–116. https://doi.org/10.1016/S2214-109X(13)70145-1.
2. Li JQ, Welchowski T, Schmid M, Mauschitz MM, Holz FG, Finger RP. Prevalence and incidence of age-related macular degeneration in Europe: a systematic review and meta-analysis. Br J Ophthalmol. 2020;104:1077–84. https://doi.org/10.1136/bjophthalmol-2019-314422.
3. Elshout M, Webers CA, van der Reis MI, de Jong-Hesse Y, Schouten JS. Tracing the natural course of visual acuity and quality of life in neovascular age-related macular degeneration: a systematic review and quality of life study. BMC Ophthalmol. 2017;17:120. https://doi.org/10.1186/s12886-017-0514-3.
4. Sivaprasad S, Bailey C, Downey L, Gilbert R, Gale R, Kotagiri A, et al. Real-world service costs for neovascular-AMD clinics in the United Kingdom: structured literature review and scenario analysis. Curr Med Res Opin. 2024;40:1221–33. https://doi.org/10.1080/03007995.2024.2362278.
5. Ho AC, Kleinman DM, Lum FC, Heier JS, Lindstrom RL, Orr SC, et al. Baseline Visual Acuity at Wet AMD Diagnosis Predicts Long-Term Vision Outcomes: An Analysis of the IRIS Registry. Ophthalmic Surg Lasers Imaging Retin. 2020;51:633–9. https://doi.org/10.3928/23258160-20201104-05.
6. Kang C, Lo JE, Zhang H, Ng SM, Lin JC, Scott IU, et al. Artificial intelligence for diagnosing exudative age-related macular degeneration - Kang, C - 2024 | Cochrane Library. https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD015522.pub2/information. Accessed October 19, 2024.
7. Krause J, Gulshan V, Rahimy E, Karth P, Widner K, Corrado GS, et al. Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy. Ophthalmology. 2018;125:1264–72. https://doi.org/10.1016/j.ophtha.2018.01.034.
8. Chen PHC, Mermel CH, Liu Y. Evaluation of artificial intelligence on a reference standard based on subjective interpretation. Lancet Digital Health. 2021;3:e693–5. https://doi.org/10.1016/S2589-7500(21)00216-8.
9. Gualino V, Tadayoni R, Cohen SY, Erginay A, Fajnkuchen F, Haouchine B, et al. Optical Coherence Tomography, Fluorescein Angiography, And Diagnosis Of Choroidal Neovascularization In Age-related Macular Degeneration. RETINA. 2019;39:1664–71. https://doi.org/10.1097/IAE.0000000000002220.
10. Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open. 2016;6:e012799. https://doi.org/10.1136/bmjopen-2016-012799.
11. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med. 2020;26:1320–4. https://doi.org/10.1038/s41591-020-1041-y.
12. Sounderajah V, Ashrafian H, Golub RM, Shetty S, De Fauw J, Hooft L, et al. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. BMJ Open. 2021;11:e047709. https://doi.org/10.1136/bmjopen-2020-047709.
13. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care. J Am Med Inform Assoc. 2020;27:2011–5. https://doi.org/10.1093/jamia/ocaa088.
14. Alderman JE, Palmer J, Laws E, McCradden MD, Ordish J, Ghassemi M, et al. Tackling algorithmic bias and promoting transparency in health datasets: the STANDING Together consensus recommendations. Lancet Digital Health. 2024;0:64. https://doi.org/10.1016/S2589-7500(24)00224-3.
15. Khan SM, Liu X, Nath S, Korot E, Faes L, Wagner SK, et al. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. Lancet Digital Health. 2021;3:e51–e66. https://doi.org/10.1016/S2589-7500(20)30240-5.
16. McCradden MD, London AJ, Gichoya JW, Sendak M, Erdman L, Stedman I et al. CANAIRI: the Collaboration for Translational Artificial Intelligence Trials in healthcare. Nat Med. 2025. https://doi.org/10.1038/s41591-024-03364-1.
17. Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. NPJ Digit Med. 2018;1:40. https://doi.org/10.1038/s41746-018-0048-y.

AUTHOR CONTRIBUTIONS

AYO – conceptualization, methodology, writing (original draft, review, editing); HDHJ – supervision, writing (review, editing); PAK – supervision, writing (review, editing).

FUNDING

AYO is supported by a National Institute for Health Research (NIHR) - Moorfields Eye Charity (MEC) Doctoral Fellowship (NIHR303691). PAK is supported by a UK Research & Innovation Future Leaders Fellowship (MR/T019050/1) and The Rubin Foundation Charitable Trust. The views expressed in this publication are those of the authors and not necessarily those of the funding organisations.

COMPETING INTERESTS

AYO, HDJH: None to declare. PAK: has acted as a consultant for Retina Consultants of America, Topcon, Roche, Boehringer-Ingleheim, and Bitfount and is an equity owner in Big Picture Medical. He has received speaker fees from Zeiss, Novartis, Gyroscope, Boehringer-Ingleheim, Apellis, Roche, Abbvie, Topcon, and Hakim Group. He has received travel support from Bayer, Topcon, and Roche. He has attended advisory boards for Topcon, Bayer, Boehringer-Ingleheim, RetinAI, and Novartis.