OPEN ACCESS

Clinical science

Validation of a deep learning system for the detection of diabetic retinopathy in Indigenous Australians

Mark A Chia ,[1,2,3] Fred Hersch,[4] Rory Sayres,[4] Pinal Bavishi,[4] Richa Tiwari,[4] Pearse A Keane ,[1,2] Angus W Turner [3,5]

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bjo-2022-322237).

[1]Institute of Ophthalmology, University College London, London, UK
[2]Moorfields Eye Hospital NHS Foundation Trust, London, UK
[3]Lions Outback Vision, Lions Eye Institute, Nedlands, Western Australia, Australia
[4]Google Health, Palo Alto, California, USA
[5]Centre for Ophthalmology and Visual Science, The University of Western Australia, Nedlands, Western Australia, Australia

Correspondence to
Dr Mark A Chia, Institute of Ophthalmology, University College London, London, EC1V 9EL, UK;
mark.a.chia@outlook.com

Received 14 July 2022
Accepted 31 December 2022
Published Online First
6 February 2023

To cite: Chia MA, Hersch F, Sayres R, *et al*. *Br J Ophthalmol* 2024;**108**:268–273.

ABSTRACT

Background/aims Deep learning systems (DLSs) for diabetic retinopathy (DR) detection show promising results but can underperform in racial and ethnic minority groups, therefore external validation within these populations is critical for health equity. This study evaluates the performance of a DLS for DR detection among Indigenous Australians, an understudied ethnic group who suffer disproportionately from DR-related blindness.

Methods We performed a retrospective external validation study comparing the performance of a DLS against a retinal specialist for the detection of more-than-mild DR (mtmDR), vision-threatening DR (vtDR) and all-cause referable DR. The validation set consisted of 1682 consecutive, single-field, macula-centred retinal photographs from 864 patients with diabetes (mean age 54.9 years, 52.4% women) at an Indigenous primary care service in Perth, Australia. Three-person adjudication by a panel of specialists served as the reference standard.

Results For mtmDR detection, sensitivity of the DLS was superior to the retina specialist (98.0% (95% CI, 96.5 to 99.4) vs 87.1% (95% CI, 83.6 to 90.6), McNemar's test p<0.001) with a small reduction in specificity (95.1% (95% CI, 93.6 to 96.4) vs 97.0% (95% CI, 95.9 to 98.0), p=0.006). For vtDR, the DLS's sensitivity was again superior to the human grader (96.2% (95% CI, 93.4 to 98.6) vs 84.4% (95% CI, 79.7 to 89.2), p<0.001) with a slight drop in specificity (95.8% (95% CI, 94.6 to 96.9) vs 97.8% (95% CI, 96.9 to 98.6), p=0.002). For all-cause referable DR, there was a substantial increase in sensitivity (93.7% (95% CI, 91.8 to 95.5) vs 74.4% (95% CI, 71.1 to 77.5), p<0.001) and a smaller reduction in specificity (91.7% (95% CI, 90.0 to 93.3) vs 96.3% (95% CI, 95.2 to 97.4), p<0.001).

Conclusion The DLS showed improved sensitivity and similar specificity compared with a retina specialist for DR detection. This demonstrates its potential to support DR screening among Indigenous Australians, an underserved population with a high burden of diabetic eye disease.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Deep learning systems (DLSs) perform well at detecting diabetic retinopathy (DR) but can underperform in racial and ethnic minority groups, therefore external validation within these populations is critical for health equity. Indigenous Australians are a disadvantaged ethnic group who suffer disproportionately from diabetic eye disease.

WHAT THIS STUDY ADDS

⇒ Compared with a retinal specialist, the DLS showed improved sensitivity and similar specificity for detecting DR in an Indigenous Australian population.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Our study supports the potential of DLSs to improve retinopathy screening in the underserved Indigenous Australian population, although further work focusing on prospective validation and real-world implementation is required.

INTRODUCTION

Diabetic retinopathy (DR) is the most common complication of diabetes and is among the leading causes of blindness in Australia.[1 2] Indigenous Australians are disproportionately affected, suffering from more than five times the rate of diabetes-related vision impairment.[3 4] Early detection and treatment through DR screening prevents vision loss in most cases, and there are clear international examples of where this has been achieved.[5] Currently, almost half of Indigenous Australians are not receiving DR screening at the frequency recommended by national guidelines,[3] in part due to insufficient availability of accessible and culturally appropriate services. With projected increases in the prevalence of diabetes, the provision of adequate DR screening services represents a major challenge for Australia.

Artificial intelligence (AI) algorithms for DR detection have shown promise in bridging the gap between demand and availability of screening resources, especially for underserved populations.[6] Deep learning, a branch of AI particularly suited to image analysis, has enabled the development of systems that can rapidly and accurately detect DR on retinal photographs,[7–15] without the need for referral to overburdened specialist services.

Despite generally performing well, an important limitation of deep learning systems (DLSs) is a tendency for reduced performance when applied to populations distinct from those in which they were developed.[16 17] These discrepancies may arise for several reasons, such as variations in normal

features or disease characteristics. Since the large training datasets required to develop a DLS tend to favour well-resourced populations, there are concerns that poor generalisability could lead to the exacerbation of healthcare inequities.[6] Furthermore, there is evidence that existing structural biases may be translated into the performance of algorithms during training.[18] Numerous examples exist within medical imaging where AI systems underperform among racial and ethnic minority groups.[19 20] Recent work has demonstrated a possible mechanism for such a bias—DLSs learn to predict racial identity even when this is unrelated to the task at hand.[21] Even more concerning, we are unable to prevent this from occurring since the basis for these predictions is unknown.[21]

The overall implication of these findings is that explicit assessment of model performance within racial and ethnic subgroups is critical.[20 21] This is particularly important for disadvantaged communities where the benefits of improved efficiency are likely to have the greatest impact. This study aims to validate a DLS for the detection of DR among Indigenous Australians, an underserved population suffering disproportionately from diabetic blindness.

MATERIALS AND METHODS

We performed a retrospective, external validation study comparing the performance of a DLS against a retina specialist for detecting DR from retinal photographs. This study follows the Standards for Reporting of Diagnostic Accuracy reporting guideline (online supplemental appendix A).[22]

Algorithm overview

Our study applied the latest Conformité Européenne-marked version of a DLS designed for DR detection (indicates conformance with European Union product legislation). The algorithm's development is described in detail by Krause *et al*.[8] In brief, a deep neural network was trained with an 'Inception-V.4' architecture to predict a 5-point DR grade, referable diabetic macular oedema (DMO), and gradability for both DR and DMO. The input to the neural network was a colour retinal photograph with a resolution of 779×779 pixels. The neural network outputs a number between 0 and 1 (indicating its confidence) for each prediction. This value is determined through multiple computational stages, parameterised by millions of numbers.

The model was trained by presenting images from a training set consisting of 2.3 million retinal photographs with a known DR severity grade. For each photograph, the model predicted its confidence for the known severity grade, slowly adjusting its parameters to improve its accuracy over time. A tuning dataset evaluated the model throughout training to determine model hyperparameters. An 'ensemble' of five individual models was then created to combine predictions for the final output. To transform the model's confidence-based outputs into discrete predictions, a threshold was used for each binary output (DMO, DR gradability and DMO gradability), and a cascade of thresholds was used to output a single DR severity level. Operating thresholds were optimised for high sensitivity suitable for a screening setting as previously described,[10] and locked prior to the commencement of this study.

Study population

This retrospective study was conducted at a single Aboriginal Community Controlled Health Service located within a metropolitan area of Perth, Western Australia. Participants were Aboriginal patients with diabetes attending a retinal screening service. Injection and laser treatment was available at monthly specialist clinics for patients identified by the screening service. The dataset consisted of retinal photographs acquired consecutively between July 2013 and October 2020. Images were non-mydriatic, single-field, 45°, macula-centred colour photographs captured using a Topcon Maestro retinal camera.

Grading and adjudication

The DLS was compared against the performance of a single human grader selected from a pool of seven United States board-certified retina specialists (mean years of postfellowship experience: 5, range: 3–10). The specialist was provided with the same colour photograph as the DLS and asked to assess gradability for DR and DMO as indicated in online supplemental appendix B. For images deemed gradable for DR, the retina specialist applied the same 5-point International Clinical Diabetic Retinopathy (ICDR) severity scale,[23] classifying images as no DR, mild non-proliferative DR (NPDR), moderate NPDR, severe NPDR or proliferative DR (PDR). For images deemed gradable for DMO, the retina specialist assessed the presence of referable DMO, defined as hard exudates within one disc diameter of the macula centre.[24] Grades were applied using an online tool-based platform that has been previously described,[25] based on disease characteristics from the ICDR severity scale. Graders were masked to the DLS and adjudication grades, and no additional clinical information was provided.

The reference standard consisted of a three-person adjudicated grade applied to all images by a panel of US board-certified retina specialists (mean 3.7 years postfellowship experience, range 1–6 years), using a method previously validated by Schaekermann *et al*.[25] In brief, each adjudicating grader first performed an independent grade using the same online platform. Images demonstrating three-person agreement were considered resolved. For unresolved cases, images were reviewed by one panel member at a time in a round-robin fashion until agreement was reached. For each review round, the active grader reviewed previous grades and comments, regraded the given image and provided further comments as required.

Outcome measures

For the primary outcomes, we combined individual assessments for gradability, DR severity and referable DMO to define the clinically relevant composite outcomes of more-than-mild DR (mtmDR), vision-threatening DR (vtDR) and all-cause referable DR. The definition of mtmDR was at least moderate NPDR or referable DMO. The definition of vtDR was at least severe NPDR or referable DMO. All-cause referable DR was defined as mtmDR or ungradable for mtmDR.

Statistical analysis

We performed sample size calculations designed for use in diagnostic accuracy studies.[26] We estimated a DLS sensitivity of 95% for detecting vtDR and set a minimum acceptable lower CI threshold of 90%. To achieve 95% confidence and 80% power, we required 183 eyes with vtDR. Assuming an ungradable rate of 15% and a vtDR prevalence of 15%,[11] this resulted in a total required sample of 1440 diabetic eyes.

Statistical analysis was performed in IBM SPSS Statistics V.26. We generated 2×2 tables to characterise the sensitivity and specificity of the DLS and retina specialist (index tests) with respect to three-person adjudication (reference standard), at the eye level. The 95% CI for sensitivities and specificities were exact Clopper-Pearson intervals and p values were calculated using

Table 1 Baseline characteristics of Indigenous Australian dataset

Characteristic	n	%
Eyes (one image per eye)	1682	
Patient demographics		
Unique individuals	864	
Mean age, years (SD)	54.9 (15.0)	
Females	453	52.4
Diabetic retinopathy grade (eyes)		
None	1091	73.6
Mild	39	2.6
Moderate	260	17.5
Severe	11	0.7
Proliferative	82	5.5
Total gradable	1483	88.2
Diabetic macular oedema grade (eyes)		
Referable diabetic macular oedema	162	11.6
Total gradable	1391	82.7

McNemar's test. Quadratic-weighted Cohen's kappa scores were calculated to measure agreement between the index tests and reference standard across the 5-point DR Scale.

RESULTS

Participants

Patient demographics and image characteristics of the external validation set are summarised in table 1. The validation set consisted of 1682 eyes of 864 patients. The mean age (SD) was 54.9 (15.0) years and women comprised 453 patients (52.4%). A flow diagram of image classification by the reference standard and DLS for mtmDR and vtDR is presented in figure 1. Of 1682 images, 1361 (80.9%) and 1348 (80.1%) images were included in the analysis for mtmDR and vtDR, respectively, with the remaining being ungradable by either the DLS, retinal specialist or reference standard.

Table 2 Comparison of deep learning system against a single retinal specialist for diabetic retinopathy detection, with reference to a three-person adjudication panel

	% (95% CI)*		
	Deep learning system	Retinal specialist	P value†
More-than-mild diabetic retinopathy‡			
Sensitivity	98.0 (96.5 to 99.4)	87.1 (83.6 to 90.6)	<0.001
Specificity	95.1 (93.6 to 96.4)	97.0 (95.9 to 98.0)	0.006
Vision-threatening diabetic retinopathy§			
Sensitivity	96.2 (93.4 to 98.6)	84.4 (79.7 to 89.2)	<0.001
Specificity	95.8 (94.6 to 96.9)	97.8 (96.9 to 98.6)	0.002
All-cause referable diabetic retinopathy¶			
Sensitivity	93.7 (91.8 to 95.5)	74.4 (71.1 to 77.5)	<0.001
Specificity	91.7 (90.0 to 93.3)	96.3 (95.2 to 97.4)	<0.001

*95% Exact Clopper-Pearson intervals.
†P value calculated between the deep learning system and retinal specialist using the McNemar test.
‡More-than-mild diabetic retinopathy (mtmDR) was defined as at least moderate non-proliferative diabetic retinopathy (NPDR) or diabetic macular oedema (DMO).
§Vision-threatening diabetic retinopathy was defined as at least severe NPDPR or DMO.
¶All-cause referable diabetic retinopathy was defined as mtmDR or ungradable for mtmDR.

Performance

Sensitivities and specificities of the DLS and retina specialist for detecting mtmDR, vtDR and all-cause referable DR are summarised in table 2. The DLS had higher sensitivity compared with the retina specialist for detection of mtmDR (98.0% vs 87.1%, p<0.001), vtDR (96.2% vs 84.4%, p<0.001) and all-cause referable DR (93.7% vs 74.4%, p<0.001). Conversely, specificity of the DLS was lower than the retina specialist;

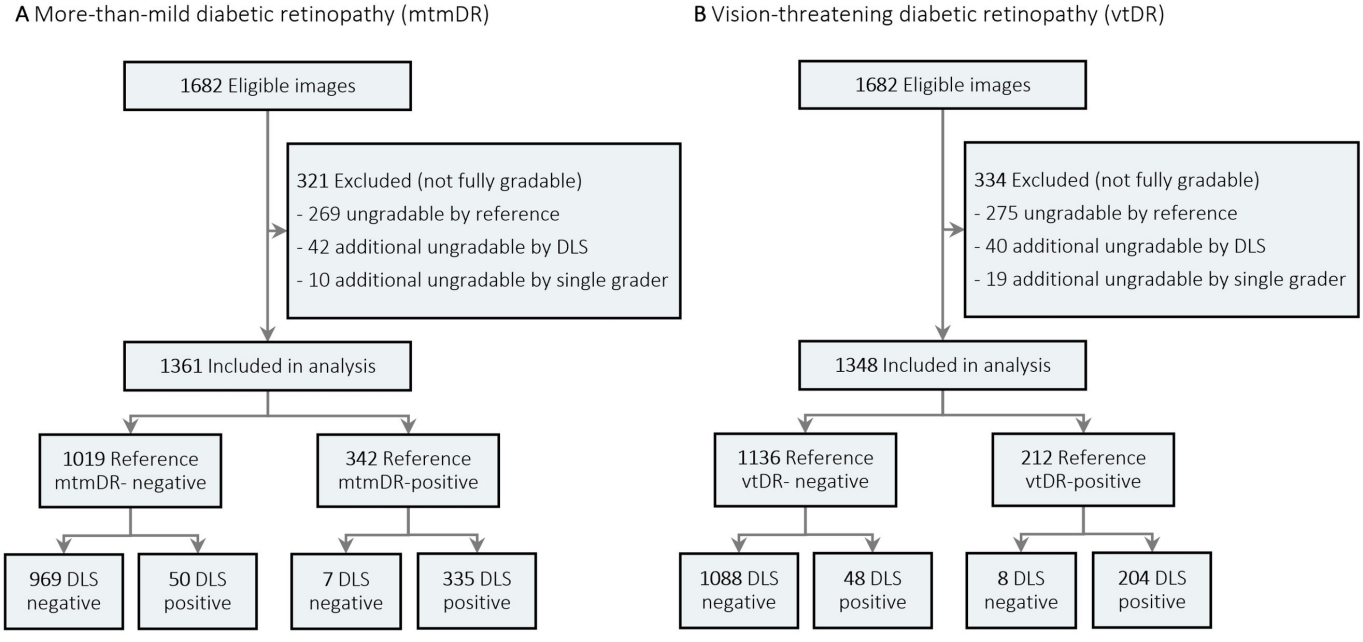


Figure 1 Flow diagram of image classification by reference standard and deep learning system (DLS). Differences in gradability arise since moderate non-proliferative diabetic retinopathy eyes that are ungradable for diabetic macular oedema are considered gradable for mtmDR but ungradable for vtDR.

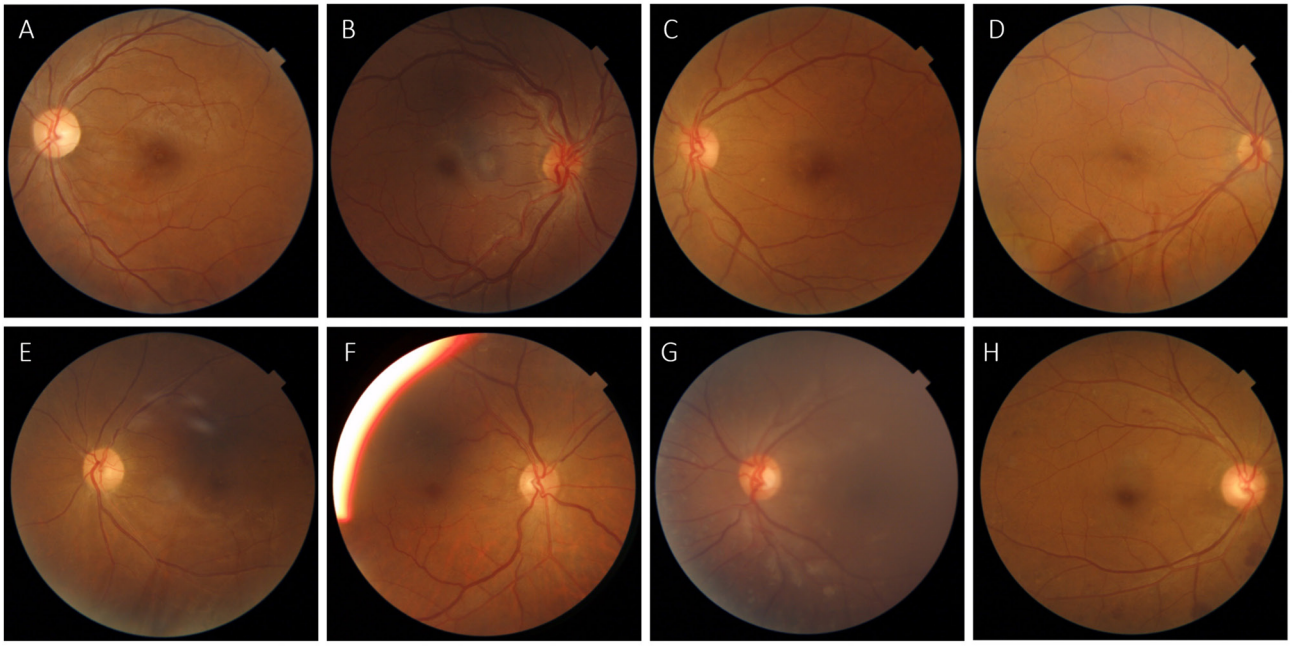


Figure 2 Retinal photographs of the 8/217 eyes diagnosed as vision-threatening diabetic retinopathy (vtDR) by the reference standard but missed by the deep learning system (DLS). According to the reference standard, A–F were graded as diabetic macular oedema (DMO), G was graded as proliferative diabetic retinopathy and H as severe non-proliferative diabetic retinopathy (NPDR). The DLS graded C as mild NPDR and the remainder as moderate NPDR, all without DMO. The single retinal specialist agreed with the DLS classification of no vtDR in all cases except D.

however, this difference was small for mtmDR (95.1 vs 97.0%, p=0.006) and vtDR (95.8% vs 97.8%, p=0.002). The reduction in specificity was larger for all-cause referable DR (91.7% vs 96.3%, p<0.001). Quadratic-weighted kappa scores for the 5-point DR Scale were not substantially different for the DLS (88.0% (95% CI, 85.5 to 90.6)) and retina specialist (89.2% (95% CI, 86.7 to 91.6)). Confusion matrices for DR severity and referable DMO are presented in online supplemental appendix C.

Gradability

The sensitivity for detecting ungradable cases of DR was higher for the DLS compared with the retina specialist (98.5% (95% CI, 96.5 to 100.0) vs 67.8% (95% CI, 61.3 to 74.4), p<0.001); however, specificity was lower (94.5% (95% CI, 93.5 to 95.8) vs 99.2% (95% CI, 99.8 to 99.6), p<0.001). For ungradable cases of DMO, the DLS showed higher sensitivity (66.7% (95% CI, 60.8 to 71.7)) vs 52.6% (95% CI, 45.7 to 57.7), p<0.001) and similar specificity (99.4% (95% CI, 98.9 to 99.8) vs 99.1% (95% CI, 98.6 to 99.6), p=0.48), although sensitivity was relatively poor for both. Confusion matrices for DR and DMO gradability are presented in online supplemental appendix C.

Misclassification analyses

The DLS missed eight cases of vtDR (false negatives) according to the reference standard. All eight retinal photographs are shown in figure 2. These misclassifications comprised six cases of missed DMO, one case of missed PDR and one case of missed severe NPDR. The DLS identified mtmDR in all but one of these instances, indicating that cases would still have been referred but with less urgency (the remaining case was graded as mild DR). In seven out of eight cases, the single retina specialist agreed with the DLS classification of no vtDR rather than the reference standard, suggesting that these were likely difficult cases. The DLS also missed seven cases of mtmDR, which were all instead graded as mild DR. The single retina specialist agreed with the DLS in four of these instances, again suggesting borderline cases.

Of 53 eyes erroneously identified by the DLS as mtmDR (false positives), the DLS identified only moderate DR (the next lowest grade) in 37 (70%) cases. Of 53 eyes erroneously identified as vtDR (false positives), the reference standard result was mtmDR and therefore still referable in 37 (70%) cases. Inspecting the 5-point DR Scale confusion matrix (online supplemental appendix C), there were 10 cases in which the DLS predicted PDR but the reference standard concluded no DR. Of these, five cases had referable pathology identified in comments by the adjudication panel (three retinal vein occlusions, two disc oedemas), and a further four had clear referable pathology identified by an ophthalmologist (AT) during post-hoc misclassification analysis (adjudicators were not specifically advised to identify non-DR pathology). The remaining case exhibited a non-referrable vascular anomaly.

DISCUSSION

Our results demonstrate that the DLS was able to identify mtmDR and vtDR with performance similar to or exceeding a retina specialist in a cohort of Indigenous Australians. For the detection of mtmDR, vtDR and all-cause referable DR, sensitivity was considerably higher than the retina specialist. Although specificity was slightly reduced for mtmDR and vtDR detection, this trade-off would likely be considered acceptable within a typical screening setting, as missed cases have the potential to lead to poor visual outcomes.

For all-cause referable DR, the reduction in specificity was larger (91.7% vs 97.5%). This remains an important consideration when evaluating the viability of a screening programme due to the cost of false positive referrals. Of the all-cause referable DR errors made by the DLS, 53% were due to misclassifications between 'no mtmDR' and 'ungradable for mtmDR,' indicating that gradability disagreements were an important source of

error. This is consistent with our findings of limited sensitivity for detecting ungradable DMO images by both the DLS (66.7%) and retina specialist (52.6%). Sensitivity for detecting ungradable images is often not consistently reported for DR detection systems.[12–15] Reviewing the confusion matrices presented in Schaekermann *et al*,[25] we noted there was poor agreement for DR gradability even between different three-person adjudication panels (mean sensitivity for detecting ungradable images was 44% across 12 comparisons). This finding implies that much of the reduction in performance for all-cause referable DR may arise due to poorly reproducible definitions of gradability, even among adjudication panels. Developing more consistent definitions of gradability may enable improved evaluation of DLS performance.

Kappa scores showed that agreement with the reference standard across the 5-point DR Scale was similar between the DLS and retina specialist. Importantly, while these scores penalise disagreements involving distant values from the reference, there is no additional penalisation for underestimating severity rather than overestimating severity. The DLS tended to overestimate severity compared with the retina specialist (online supplemental appendix C), which is generally a more acceptable error in a screening context. Misclassification analyses illustrated that DLS errors usually occurred in difficult or borderline cases. In most cases, these errors involved a misclassification to the adjacent category in the severity scale. Only eight cases of vtDR were missed and the single retina specialist agreed with the DLS in all but one of these instances.

This DLS has previously been applied to external validations sets in India[10] and Thailand[9] and results from our novel population group were comparable. For detecting moderate or worse DR in these studies, point estimates ranged between 88.9% and 96.8% for sensitivity and 92.2% and 95.6% for specificity. Reported performance for other DLSs for referable DR detection have ranged between 87.2% and 97.5% for sensitivity and 87.0% and 98.5% for specificity; however, definitions, study populations and methodology vary considerably.[16]

Our study has several strengths. First, the DLS was evaluated in a novel population suffering from a high burden of diabetic-eye disease. Second, classification thresholds were locked prior to the commencement of the study rather than being derived through post-hoc analysis of receiver operating curves. Third, we applied a consistent, rigorous reference standard to all images for external validation. Fourth, we report a range of composite outcomes that are clinically relevant to real-world screening programmes, including all-cause referable DR.

Our study has relevant limitations. Despite the use of a rigorous reference standard, we did not use optical coherence tomography imaging to define the presence of DMO, as has been recently described.[27] The reference standard also did not include identification of non-DR referrable pathology. Although the DLS did identify important non-DR pathology in our misclassification analysis, it is possible that there was additional pathology that a retina specialist would have detected beyond the DLS. Our data came from a single centre, therefore our findings may not generalise to other Indigenous populations or to settings using alternative screening strategies such as multifield or dilated photography. Finally, as a retrospective study our validation set may not reflect the disease spectrum and challenges of a prospective cohort.

Future work should aim to address several challenges which remain for DLS-driven DR screening, with a focus on prospective validation and real-world implementation. Given the costs associated with false positive referrals using a fully automated model, the development of a hybrid model may provide a more practical option for implementation.[28] This would involve the use of a DLS to rule out non-referable cases followed by secondary human assessment.

Careful consideration of processes for integrating DLSs into clinical-care pathways is critical, especially for Indigenous Australians. In addition to lower screening rates, Indigenous patients experience reduced follow-up after referral.[29] Proposed explanations for this include: (1) higher proportions living in areas serviced by visiting specialists, (2) reduced accessibility through conventional communication pathways such as mail and telephone and (3) poor understanding of the need for attendance.[29] A key benefit of a DLS is the ability to provide an immediate referral decision at the time of screening, facilitating in-person education and appointment planning. Although there is some supporting evidence derived from other settings that such a pathway would result in increased referral adherence,[30 31] further work in this area is needed.

Prospective validation studies to date have identified relevant implementation challenges including poor internet availability and technical issues limiting consistent acquisition of gradable photographs.[32] Large-scale deployment of a DLS for retinal screening is dependent on addressing these difficulties with validated solutions. In addition, it is known that a range of complex cultural factors influence the acceptability and uptake of healthcare interventions for Indigenous Australians, therefore collaboration with community leaders is essential.[29] Fear and distrust towards Western medical practices is an important barrier to healthcare access in Indigenous communities, and it is possible that similar concerns may limit the uptake of AI-based solutions.

Our study shows that a DLS can detect DR in an Indigenous Australian cohort with improved sensitivity and similar specificity compared with a retina specialist. This demonstrates the potential of the system to support DR screening among Indigenous Australians, an underserved population with a high burden of diabetic eye disease. Inadequate DR screening represents an important source of healthcare inequity and is therefore an urgent priority for Australia.

Twitter Mark A Chia @markachia and Pearse A Keane @pearsekeane

Acknowledgements The authors thank Yun Liu and Naama Hammel for manuscript review; Derek Wu, Roy Lee and the labelling software team in Google Health for assistance in data labelling; and Derbarl Yerrigan Health Service for technical and logistical support. Part of this work was presented as an abstract at the Association for Research in Vision and Ophthalmology (ARVO) Annual Meeting 2022.

Contributors MAC: research design, data acquisition, data analysis, data interpretation, manuscript preparation, and guarantor. FH: research design, data interpretation and manuscript revision. RS: data analysis, data interpretation and manuscript revision. PB, RT and PAK: data interpretation and manuscript revision. AT: research design, data interpretation and manuscript revision. All authors approved the final manuscript.

Funding Google LLC funded this study, and participated in the design of the study, conducting the study, data collection, data management, data analysis, interpretation of the data, preparation, review and approval of the manuscript. MAC: Supported by a General Sir John Monash Scholarship. PAK: Supported by a Moorfields Eye Charity Career Development Award (R190028A) and a UK Research & Innovation Future Leaders Fellowship (MR/T019050/1).

Competing interests PAK has acted as a consultant for DeepMind, Roche, Novartis and Apellis and is an equity owner in Big Picture Medical. He has received speaker fees from Heidelberg Engineering, Topcon, Allergan and Bayer. FH, RS, PB and RT are employees of Google LLC and own Alphabet stock.

Patient consent for publication Not applicable.

Ethics approval This study involves human participants and was approved by Western Australian Aboriginal Health Ethics Committee (Reference Number: 864). The requirement for informed consent was waived due to the retrospective nature of the study and the use of fully anonymised retinal photographs.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

ORCID iDs
Mark A Chia http://orcid.org/0000-0003-0339-5186
Pearse A Keane http://orcid.org/0000-0002-9239-745X
Angus W Turner http://orcid.org/0000-0001-5949-7451

REFERENCES

1 Simó-Servat O, Hernández C, Simó R. Diabetic retinopathy in the context of patients with diabetes. *Ophthalmic Res* 2019;62:211–7.
2 Heath Jeffery RC, Mukhtar SA, McAllister IL, *et al*. Inherited retinal diseases are the most common cause of blindness in the working-age population in Australia. *Ophthalmic Genet* 2021;42:431–9.
3 Foreman J, Keel S, Xie J, *et al*. National eye health survey report. Melbourne: Centre for Eye Research Australia; 2016. Available: https://www.vision2020australia.org.au/wp-content/uploads/2019/06/National-Eye-Health-Survey_Full-Report_FINAL.pdf
4 Chia MA, Taylor JR, Stuart KV, *et al*. Prevalence of diabetic retinopathy in indigenous and non-indigenous australians: a systematic review and meta-analysis. *Ophthalmology* 2023;130:56–67.
5 Liew G, Michaelides M, Bunce C. A comparison of the causes of blindness certifications in England and Wales in working age adults (16-64 years), 1999-2000 with 2009-2010. *BMJ Open* 2014;4:e004015.
6 Ibrahim H, Liu X, Zariffa N, *et al*. Health data poverty: an assailable barrier to equitable digital health care. *Lancet Digit Health* 2021;3:e260–5.
7 Gulshan V, Peng L, Coram M, *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus Photographs. *JAMA* 2016;316:2402–10.
8 Krause J, Gulshan V, Rahimy E, *et al*. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology* 2018;125:1264–72.
9 Raumviboonsuk P, Krause J, Chotcomwongse P, *et al*. Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. *NPJ Digit Med* 2019;2:25.
10 Gulshan V, Rajan RP, Widner K, *et al*. Performance of a deep-learning algorithm vs manual grading for detecting diabetic retinopathy in India. *JAMA Ophthalmol* 2019;137:987–93.
11 Scheetz J, Koca D, McGuinness M, *et al*. Real-world artificial intelligence-based opportunistic screening for diabetic retinopathy in endocrinology and Indigenous healthcare settings in Australia. *Sci Rep* 2021;11:15808.
12 Ipp E, Liljenquist D, Bode B, *et al*. Pivotal evaluation of an artificial intelligence system for autonomous detection of referrable and vision-threatening diabetic retinopathy. *JAMA Netw Open* 2021;4:e2134254.
13 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.
14 Ting DSW, Cheung CY-L, Lim G, *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* 2017;318:2211–23.
15 Li Z, Keel S, Liu C, *et al*. An automated grading system for detection of vision-threatening referable diabetic retinopathy on the basis of color fundus Photographs. *Diabetes Care* 2018;41:2509–16.
16 Ting DSW, Peng L, Varadarajan AV, *et al*. Deep learning in ophthalmology: the technical and clinical considerations. *Prog Retin Eye Res* 2019;72:100759.
17 Chia MA, Turner AW. Benefits of integrating telemedicine and artificial intelligence into outreach eye care: stepwise approach and future directions. *Front Med (Lausanne)* 2022;9:835804.
18 Obermeyer Z, Powers B, Vogeli C, *et al*. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019;366:447–53.
19 Pierson E, Cutler DM, Leskovec J, *et al*. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nat Med* 2021;27:136–40.
20 Seyyed-Kalantari L, Liu G, McDermott M, *et al*. CheXclusion: fairness gaps in deep chest X-ray classifiers. *arXiv* 2020.
21 Banerjee I, Bhimireddy AR, Burns JL, *et al*. Reading race: AI recognises patient's racial identity in medical images. *arXiv* 2021.
22 Bossuyt PM, Reitsma JB, Bruns DE, *et al*. Stard 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ* 2015;351:h5527.
23 Wilkinson CP, Ferris FL 3rd, Klein RE, *et al*. Proposed International clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology* 2003;110:1677–82.
24 Bresnick GH, Mukamel DB, Dickinson JC, *et al*. A screening approach to the surveillance of patients with diabetes for the presence of vision-threatening retinopathy. *Ophthalmology* 2000;107:19–24.
25 Schaekermann M, Hammel N, Terry M, *et al*. Remote tool-based adjudication for grading diabetic retinopathy. *Transl Vis Sci Technol* 2019;8:40.
26 Flahault A, Cadilhac M, Thomas G. Sample size calculation should be performed for design accuracy in diagnostic test studies. *J Clin Epidemiol* 2005;58:859–62.
27 Liu X, Ali TK, Singh P, *et al*. Deep learning to detect OCT-derived diabetic macular edema from color retinal photographs: a multicenter validation study. *Ophthalmol Retina* 2022;6:398–410.
28 Xie Y, Nguyen QD, Hamzah H, *et al*. Artificial intelligence for teleophthalmology-based diabetic retinopathy screening in a national programme: an economic analysis modelling study. *Lancet Digit Health* 2020;2:e240–9.
29 Copeland S, Muir J, Turner A. Understanding Indigenous patient attendance: a qualitative study. *Aust J Rural Health* 2017;25:268–74.
30 Pedersen ER, Cuadros J, Khan M, *et al*. Redesigning clinical pathways for immediate diabetic retinopathy screening results. *NEJM Catalyst* 2021;2.
31 Mathenge W, Whitestone N, Nkurikiye J, *et al*. Impact of artificial intelligence assessment of diabetic retinopathy on referral service uptake in a low-resource setting: the RAIDERS randomized trial. *Ophthalmol Sci* 2022;2:100168.
32 Ruamviboonsuk P, Tiwari R, Sayres R, *et al*. Real-time diabetic retinopathy screening by deep learning in a multisite national screening programme: a prospective interventional cohort study. *Lancet Digit Health* 2022;4:e235–44.

Online-only Supplementary Material

Chia MA, *et al. Br J Ophthalmol* 2023;0:1–6. doi: 10.1136/bjo-2022-322237

Appendix A: STARD Checklist

Section & Topic	No	Item	Reported page #
TITLE OR ABSTRACT			
	1	Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)	3
ABSTRACT			
	2	Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)	3
INTRODUCTION			
	3	Scientific and clinical background, including the intended use and clinical role of the index test	6
	4	Study objectives and hypotheses	6
METHODS			
Study design	**5**	Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)	8
Participants	**6**	Eligibility criteria	9
	7	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)	9
	8	Where and when potentially eligible participants were identified (setting, location and dates)	9
	9	Whether participants formed a consecutive, random or convenience series	9
Test methods	**10a**	Index test, in sufficient detail to allow replication	9
	10b	Reference standard, in sufficient detail to allow replication	9
	11	Rationale for choosing the reference standard (if alternatives exist)	9
	12a	Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory	9
	12b	Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory	N/A
	13a	Whether clinical information and reference standard results were available to the performers/readers of the index test	N/A
	13b	Whether clinical information and index test results were available to the assessors of the reference standard	9
Analysis	**14**	Methods for estimating or comparing measures of diagnostic accuracy	10
	15	How indeterminate index test or reference standard results were handled	10
	16	How missing data on the index test and reference standard were handled	10
	17	Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory	N/A
	18	Intended sample size and how it was determined	10
RESULTS			
Participants	**19**	Flow of participants, using a diagram	Figure 1
	20	Baseline demographic and clinical characteristics of participants	Table 1
	21a	Distribution of severity of disease in those with the target condition	Table 1
	21b	Distribution of alternative diagnoses in those without the target condition	N/A
	22	Time interval and any clinical interventions between index test and reference standard	N/A
Test results	**23**	Cross tabulation of the index test results (or their distribution) by the results of the reference standard	Appendix
	24	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	Table 2
	25	Any adverse events from performing the index test or the reference standard	N/A
DISCUSSION			
	26	Study limitations, including sources of potential bias, statistical uncertainty, and generalisability	16
	27	Implications for practice, including the intended use and clinical role of the index test	17
OTHER INFORMATION			
	28	Registration number and name of registry	N/A
	29	Where the full study protocol can be accessed	N/A
	30	Sources of funding and other support; role of funders	18

Chia MA, *et al. Br J Ophthalmol* 2023;0:1–6. doi: 10.1136/bjo-2022-322237

Appendix B: Gradability Instructions and Examples

DR Gradability

How gradable is the image for DR? Note: This question doesn't show if "Other" was selected as Fundus field. Also, "Gradable" and "Gradable with Difficulty" were both considered as gradable images.

Gradable	• You can clearly see the features of DR in regions you'd expect to see in a given fundus field. This does not mean that you can confidently make a full diagnosis for DR with just this image.
Gradable with Difficulty	• Images show key regions for the defined field of view, but image quality is not good enough to allow for a confident grading • Some key regions may be blurry or missing, but clearly visible regions show obvious pathology/features which point to at least moderate DR • If visible regions don't show any pathology, then the image is "ungradable" as below
Ungradable	• Images don't show key regions with good enough quality for a confident grading. Also the other visible areas do not show any obvious pathology

DME Gradability

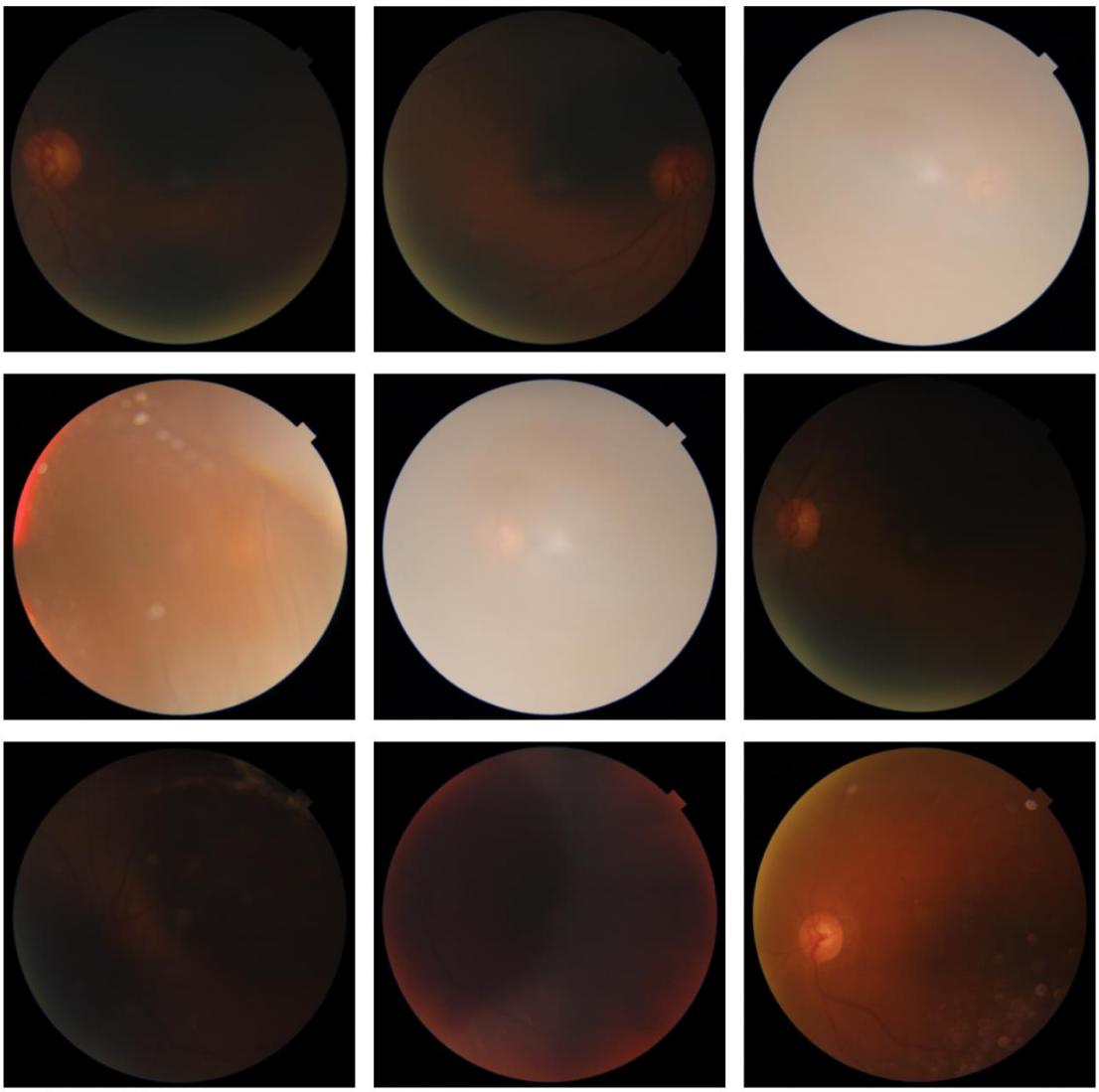
How gradable is the image for DME? Note: This question doesn't show if "Other" was selected as the Fundus field. Also, "Gradable" and "Gradable with Difficulty" were both considered as gradable images.

Gradable	• Entire macula (one disc diameter from fovea center) can be seen clearly. A confident diagnosis can be made.
Gradable with Difficulty	• Entire macula can be seen, but the image quality is not good enough to make a confident diagnosis. • Part of the macula is missing, but there's strong evidence of DME (e.g hard exudates) in the visible area
Ungradable	• Macula is not visible or only partially visible (less than one disc diameter from fovea center) either because it is not in the field of view or because it is occluded by artifacts, dark shadow etc. What can be seen is not enough to rule out DME. • DR symptoms are not clear in the image (ungradable for any symptom of DR) then mark image as ungradable for DME even though hard exudates are visible. (As hard exudates may not be due to Diabetic macular Edema.

Chia MA, *et al. Br J Ophthalmol* 2023;0:1–6. doi: 10.1136/bjo-2022-322237

Examples

Images that were ungradable for diabetic retinopathy by the deep learning system and adjudicators.



Appendix C: Confusion Matrices

DLS vs Reference - DR Confusion Matrix

		DLS						
		Ungradable	No DR	Mild NPDR	Mod NPDR	Severe NPDR	PDR	Total
Reference	Ungradable	196	1	1	0	0	1	199
	No DR	73	914	66	28	0	10	1091
	Mild NPDR	0	1	25	13	0	0	39
	Mod NPDR	6	1	7	150	70	26	260
	Severe NPDR	0	0	0	1	7	3	11
	PDR	2	0	0	1	0	79	82
Total		277	917	99	193	77	119	1682

Retina specialist vs Reference - DR Confusion Matrix

		Retina specialist						
		Ungradable	No DR	Mild NPDR	Mod NPDR	Severe NPDR	PDR	Total
Reference	Ungradable	135	51	1	11	0	1	199
	No DR	5	1047	18	20	1	0	1091
	Mild NPDR	1	13	22	3	0	0	39
	Mod NPDR	4	27	9	207	10	3	260
	Severe NPDR	0	0	0	8	3	0	11
	PDR	2	5	1	10	1	63	82
Total		147	1143	51	259	15	67	1682

DLS vs Reference - DME Confusion Matrix

		DLS			
		Ungradable	No DME	DME	Total
Reference	Ungradable	194	82	15	291
	No DME	7	1154	68	1229
	DME	1	7	154	162
Total		202	1243	237	1682

Retina specialist vs Reference - DME Confusion Matrix

		Retina specialist			
		Ungradable	No DME	DME	Total
Reference	Ungradable	153	130	8	291
	No DME	8	1192	29	1229
	DME	4	24	134	162
Total		165	1346	171	1682

DLS vs Reference - mtmDR Confusion Matrix

		DLS			
		Ungradable	No mtmDR	mtmDR	Total
Reference	Ungradable	231	32	6	269
	No mtmDR	35	972	53	1060
	mtmDR	7	7	339	353
Total		273	1011	398	1682

Retina specialist vs Reference - mtmDR Confusion Matrix

		Retina specialist			
		Ungradable	No mtmDR	mtmDR	Total
Reference	Ungradable	137	118	14	269
	No mtmDR	4	1034	22	1060
	mtmDR	6	42	305	353
Total		147	1194	341	1682

DLS vs Reference - vtDR Confusion Matrix

		DLS			
		Ungradable	No vtDR	vtDR	Total
Reference	Ungradable	233	35	7	275
	No vtDR	36	1097	53	1186
	vtDR	4	8	209	221
Total		273	1140	269	1682

Retina specialist vs Reference - vtDR Confusion Matrix

		Retina specialist			
		Ungradable	No vtDR	vtDR	Total
Reference	Ungradable	141	125	9	275
	No vtDR	7	1160	19	1186
	vtDR	6	34	181	221
Total		154	1319	209	1682

DLS vs Reference – All-cause referable DR

Count

		DLS		
		Non-referable	Referable	Total
Reference	Non-referable	972	88	1060
	Referable	39	583	622
Total		1011	671	1682

Retina specialist vs Reference – All-cause referable DR

Count

		Retina specialist		
		Non-referable	Referable	Total
Reference	Non-referable	1034	26	1060
	Referable	160	462	622
Total		1194	488	1682

Note: Slight discrepancies exist when comparing matrices with Figure 1 and Table 2. Discrepancies arise due to the need to exclude images ungradable by any one of the reference, DLS, or retina specialist when performing pairwise statistical testing.

Chia MA, *et al. Br J Ophthalmol* 2023;0:1–6. doi: 10.1136/bjo-2022-322237