

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:

Quinn, N;Brazionis, L;Zhu, B;Ryan, C;D'Aloisio, R;Lilian Tang, H;Peto, T;Jenkins, A

Title:

Facilitating diabetic retinopathy screening using automated retinal image analysis in underresourced settings

Date:

2021-09-01

Citation:

Quinn, N., Brazionis, L., Zhu, B., Ryan, C., D'Aloisio, R., Lilian Tang, H., Peto, T. & Jenkins, A. (2021). Facilitating diabetic retinopathy screening using automated retinal image analysis in underresourced settings. *Diabetic Medicine*, 38 (9), <https://doi.org/10.1111/dme.14582>.

Persistent Link:

<https://hdl.handle.net/11343/298457>

DR NICOLA QUINN (Orcid ID : 0000-0002-3224-5091)

DR LAIMA BRAZIONIS (Orcid ID : 0000-0002-5740-9662)

Article type : Research Article

Target Journal: Diabetic Medicine

Title: Facilitating diabetic retinopathy screening using automated retinal image analysis in under-resourced settings

Running Title: Automated image analysis in diabetic retinopathy screening

Authors: Nicola Quinn^{1,3}, Laima Brazionis^{1,2}, Benjamin Zhu¹, Chris Ryan^{1,2}, Rossella D'Aloisio^{3,4}, Hongying Lilian Tang^{5*}, Tunde Peto^{3*}, Alicia Jenkins^{1-3*} on behalf of the Centre of Research Excellence in Diabetic Retinopathy Study and TEAMSnet Study Groups¹

^ Equal first authors, * equal senior authors

Institutions:

1. NHMRC Clinical Trials Centre, The University of Sydney, NSW; Australia
2. Department of Medicine, The University of Melbourne, VIC; Australia
3. Centre for Public Health, Queen's University Belfast, UK
4. Department of Medicine and Science of Ageing, Ophthalmology Clinic, University "G. d'Annunzio" Chieti-Pescara, Chieti; Italy
5. Department of Computer Science, University of Surrey, UK

Corresponding author

Professor Tunde Peto
Institute of Clinical Sciences, Block A,
Centre for Public Health,
Queen's University Belfast

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the [Version of Record](#). Please cite this article as [doi: 10.1111/DME.14582](https://doi.org/10.1111/DME.14582)

This article is protected by copyright. All rights reserved

Belfast, Northern Ireland BT12 6BA

Phone: +44(0)28 90976400

Email: t.peto@qub.ac.uk

Abstract words: 250

Manuscript words: 2935

Conflicts of interest: No conflicts of interest to declare

What's new?

- Indigenous Australians are almost four times more likely to have diabetes than non-Indigenous Australians, with diabetic retinopathy (DR) screening coverage much lower in remote and very remote Australia.
- Automated retinal image analysis (ARIA) software has not been applied to Indigenous Australians previously and in this study it is effective in triaging the presence and absence of DR.
- ARIA may facilitate DR screening programmes in these high-risk populations with retinal cameras for use in DR screening in non-urban regions of Australia and by also substantially reducing the workload of human graders.

Acknowledgements

The authors wish to thank the TEAMSnet Study Group, Partners and Collaborators:
TEAMSnet Study Group:

Chief investigators: Professor Sven-Erik Bursell, Professor Alex Brown, Professor Alicia Jenkins, Dr David O'Neal, Professor Danny Liew

Associate Investigators: Professor Tien Wong, Professor Hugh Taylor, Professor Anthony Keech, Professor Kerin O'Dea, Professor Ecosse Lamoureux, Dr Mark Horton, TEAMSnet Programme

Director: Dr Laima Brazionis, TEAMSnet Project

Manager: Christopher Ryan

Partners: Wurli Wurlijang (Katherine); Miwatj (Nhulunbuy); Central Australian Aboriginal Congress (Alice Springs); Aboriginal Medical Services Alliance (NT); CERA (Melbourne); Estenda Solutions (USA); the Fred Hollows Foundation (Global); The University of Melbourne; NHMRC Clinical Trials Centre (University of Sydney).

Abstract

Aim

To evaluate an automated retinal image analysis (ARIA) of indigenous retinal fundus images against a human grading comparator for the classification of diabetic retinopathy (DR) status.

Methods

Indigenous Australian adults with type 2 diabetes (n=410) from three remote and very-remote primary-care services in the Northern Territory, Australia, underwent tele-retinal DR screening. A single central retinal fundus photograph (opportunistic mydriasis) for each eye was later re-graded using a single ARIA and a UK human grader and national DR classification system. The sensitivity and specificity of ARIA were assessed relative to the comparator. Proportionate agreement and a Kappa statistic were also computed.

Results

Retinal images from 391 and 393 participants were gradable for 'Any DR' by the human grader and ARIA grader, respectively. 'Any DR' was detected by the human grader in 185 (47.3%) participants and by ARIA in 202 (48.6%) participants (agreement=88.0%, Kappa =0.76,), while proliferative DR was detected in 31 (7.9%) and 37 (9.4%) participants (agreement=98.2%, Kappa=0.89,), respectively. The ARIA software had 91.4 (95% CI 86.3-95.0) sensitivity and 85.0 (95% CI 79.3-89.5) specificity for detecting 'Any DR' and 96.8 (95% CI 83.3-99.9) sensitivity and 98.3 (95% CI 96.4-99.4) specificity for detecting proliferative DR.

Conclusions

This ARIA software has high sensitivity for detecting 'Any DR' hence could be used as a triage tool for human graders. High sensitivity was also found for detection of proliferative DR by ARIA. Future versions of this ARIA should include maculopathy and referable DR (CSME and / or PDR). Such ARIA software may benefit diabetes care in less-resourced regions.

Key words: diabetic retinopathy, Indigenous Australians, automated retinal image analysis

Author Manuscript

81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114

Introduction

Diabetic retinopathy (DR) is a major cause of vision loss globally (1). Detection through clinical examination or grading of retinal fundus images is essential to identify sight-threatening retinopathy (STDR) to initiate timely ocular treatment and prevent vision loss, in addition to signalling need for control of systemic risk factors such as smoking and hyperglycaemia, which also attenuates the risk of other chronic diabetes complications (2). Detection of early DR is important to facilitate DR risk factor control and modify DR screening frequency. Annual DR screening is generally recommended in Australia, however the global diabetes prevalence is expected to increase from 463 million in 2019 to 700 million by 2045 (3) therefore DR screening faces major challenges.

National DR screening programmes, including that of the UK National Health Service Diabetic Eye Screening (DES) Programme, are effective, however require considerable time analysing images by trained, certified and quality assured human graders (4). DR screening guidelines acknowledge the need for innovative DES programmes which will vary depending on the community and resources available. People in less advantaged areas often have competing priorities, poorer access to care, time constraints, cost barriers, insufficient health literacy and awareness of DR, and often a lack of care coordination (5). Due to the expected increase in imaging volume, new solutions such as automated retinal image analysis (ARIA) may be an option to successfully analyse large volumes of retinal images on-site. In contrast to manual off-site human grading, ARIA may facilitate real-time point-of-care DR grading.

ARIA has been extensively reviewed, with some systems reporting high sensitivity and specificity in DR screening (6, 7). ARIA may reduce the need for manual graders and cost by decreasing the number of images requiring manual grading (8). Early ARIA programs were designed to detect specific DR characteristics such as microaneurysms or exudates (9-14), subsequently developing to examining the presence or absence of any retinopathy. More recently ARIA developers began testing whether ARIA could differentiate between DR

severity levels (15-17). Although ARIA performs well in retinal screening facilities using photographs taken by experienced photographers, it does not obviate an individual's need for comprehensive eye exams based on their ocular and family history. There is limited research into ARIA's accuracy in less advantaged populations or low resource settings using photos taken by non-ophthalmic clinicians (non-ophthalmologists and non-optometrists, including locally trained Indigenous workers, nurses and locum general practitioners) and where there is often high clinic and study staff turnover. To be clinically applicable, it is essential that ARIA performs well in areas with less experienced healthcare workers or non-ophthalmic clinicians. Therefore, this study aims to compare the accuracy of ARIA DR grading to manual DR grading [comparator], based on retinal images of Indigenous Australians with Type 2 diabetes acquired by locally-trained health workers in remote primary care clinical settings.

Participants and Methods

Study Design and Participants

The main aim was to assess the grading performance and diagnostic accuracy of an ARIA, using the UK National Health Service DES Programme manual grading as the reference standard. The Telehealth Eye and Associated Medical Services Network (TEAMSnet) study design and protocol have been previously published (18). The study was approved by the University of Melbourne, the Northern Territory Menzies School of Health Research and the Central Australian Human Research Ethics Committees. All participants provided written informed consent prior to participation.

Indigenous Australians (age $\geq$ 18 years) from the Northern Territory of Australia with diagnosed Type 2 diabetes were recruited (11/2013-12/2015) from three remote primary-care services using active and opportunistic recruitment strategies. In Australia geographical accessibility to goods and services is reflected by the Accessibility/Remoteness Index of Australia, which measures the remoteness of a point based on the physical road distance to the nearest Urban Centre in each of five size classes. The enhanced preferred version currently in use is the Australian Standard Geographical Classification (ASGC) Remoteness Area classification of six Remoteness Area classes - Major Cities, Inner Regional, Outer Regional, Remote, Very Remote and Migratory (19).

DR screening included presenting (aided or unaided) visual acuity (VA) (Snellen VA chart). VA was categorised as; normal (6/6 or better), reduced ($<6/6$ but no worse than 6/12), impaired ($<6/12$ but no worse than 6/60) and blind ($<6/60$)(20). Pupils were dilated with one drop each of 1% tropicamide and 2.5% phenylephrine if indicated and clinical supervision was available. If undilated pupils were $\geq 4\text{mm}$ or mydriatic drops were contraindicated, dilation was not attempted. Brazionis *et al* reported opportunistic dilation in 54% (n=162) participants in the Central Australian diabetic retinopathy screening study, thought to have been similar to the current study (21). Using a Canon CR-2 fundus camera a central retinal image was captured by locally trained staff. Effectiveness of care may decrease with high staff turnover and use of short-term staff (22), as herein. Images, automatically captured by the Global Retinal Imager software application, were securely transmitted as converted (DICOM) HL7 messages over a Secure Socket Layer encrypted tunnel to the Melbourne Centre for Eye Research Australia (CERA) server. De-identified retinal images were then transferred in person (LB) using a portable hard-drive to Queen's University Belfast, UK for manual grading and to the University of Surrey, UK for automatic retinal grading.

Human Retinal Grading

Images were graded by a single trained and certified grader at the Belfast Ophthalmic Reading Centre (BORC). Grading was based on the National Health Service DES Programme (4). Each participant was categorised into no apparent DR (R0), background retinopathy (R1), pre-proliferative retinopathy (R2) or proliferative retinopathy (R3). Any ungradable images were listed as such. If the grader was unsure of the grade, images were adjudicated by an expert (ophthalmologist) clinician (TP).

Automated Retinal Image Analyses System

The ARIA system has been described in detail and validated previously (23).

Discordant cases

In cases where ARIA disagreed with the human grading original images were reviewed by a research fellow (NQ). Ten percent of images checked by NQ were also graded by an ophthalmic clinician and retinal grading specialist (LB).

Statistical Analysis

Data were analysed using IBM SPSS Statistics for Windows, Version 25.0 (Armonk, NY:

IVM Corp). DR prevalence was determined for both human and ARIA grading. Cross-tabulation was used to compute DR stages between ARIA and the human grader. Differences between the human grader and ARIA were assessed using sensitivity and specificity analysis. With the human grader as the reference standard and ARIA as the test grading protocol, we computed sensitivity and specificity values for the ability of ARIA to a) diagnose any retinopathy (R1-R3) and b) differentiate between DR stages. DR prevalence was reported on participant-level (clinical) data, and all other outcomes were reported on eye-level data. While results at eye level may be more relevant to assess raw algorithm performance, those at participant level are more clinically relevant. Proportionate agreement and Kappa statistic was computed between ARIA and human grading for participants with Any DR and DR severity stages.

Results

Of 410 Indigenous Australians screened, 400 had retinal images for both eyes and 10 participants had retinal images for a single eye. For ARIA and human grading 60 eyes (43 participants) and 68 eyes (50 participants) were ungradable, respectively. For human grading 742 eyes from 391 participants were available for analysis. For ARIA grading 750 eyes from 393 participants were available for analysis. A total of 391 participants were graded by both human and ARIA. Reasons for ungradable ARIA images included; cataract (21.7%), poor image quality (71.7%), unknown (5.0%) and pterygium (1.7%). Human grading had an additional eight ungradable images due to cataracts and poor image quality.

As shown in Table 1A, DR prevalence, based on human grading was 47.3% (n=185), while 116 (29.7%), 38 (9.7%) and 31 (7.9%) participants were classified as R1, R2 or R3, respectively. Using ARIA grading, overall DR prevalence was 48.6% (n=202) with 110 (28.0%), 55 (14.0%) and 37 (9.4%) classified as R1, R2 or R3, respectively.

Table 1A also shows discordant cases between ARIA and human grading. A total of 57/391 (14.6%) cases were deemed false negatives and 42/391 (10.7%) were considered false positives for R1-R3. After arbitration, the main causes for false positives included poor image quality / small pupil 10/42 (23.8%), cataract 7/42 (16.7%), artefacts 6/42 (14.3%) and other pathology 2/42 (4.8%). The main reasons for false negatives included other pathology / treatment 11/57 (19.3%), cataract 17/57 (29.8%), poor image quality / small

pupil 14/57 (24.6%) and artefacts 11/57 (19.3%).

Comparing ARIA to the reference standard (human grader), the sensitivity and specificity of ARIA software to diagnose prevalent DR [per eye] was 91.7% (95% CI, 88.0% - 94.6%) and 90.5% (95% CI, 87.3% - 93.0%), respectively. Sensitivities and specificities (%) for each individual retinopathy grade ranged between 74.6 - 95.7% and 90.5 - 98.6%, respectively (Table 2B). For participant level data, ARIA software sensitivity and specificity to diagnose any DR was 91.4% (95% CI, 86.3 - 95.0) and 85.0% (95% CI, 79.3 - 89.5), respectively. For R1-R3 DR levels, sensitivity and specificity of ARIA ranged 70.7% - 96.8% and 89.8% - 98.3%, respectively, with R1 having the lowest sensitivities for both participants and all available eyes (Table 2B). Agreement analysis between ARIA and the human grader for 'Any DR' and each retinopathy stage ranged 84.1% - 98.2%. Proportionate agreement and kappa values were 88.0% (0.76), 84.1% (0.61), 92.3% (0.63), 98.2% (0.89) for 'Any DR', R1, R2 and R3 respectively.

Table 2 shows details of ARIA and human grading where a) both eyes had the same grade, b) only one eye had the same grade or c) both eyes had different grades. In 317 (77.3%) participants both eyes were graded the same by ARIA and the human grader, 57 (13.9%) participants had one eye the same grade and in 19 (6.3%) participants both eyes had different DR grades. Ten participants (2.5%) had images for one eye only. Of these, eight had the same grade in ARIA and human grading. ARIA graded worse than the human (human=R1, ARIA=R2) in one eye and the 10th eye was ungradable by the human grader whereas ARIA graded that eye as R3. Supplementary Table 1 compares ARIA and human grading for all eyes. ARIA tended to grade a higher level of DR than the human grader, which is commented on in the discussion.

Discussion

To our knowledge, this is the first study to utilise ARIA software in Indigenous Australians with diabetes. Study results show that the ARIA software may assist the detection of 'Any DR', in triaging images that need to be manually graded. This study used a single ARIA system by the University of Surrey and other ARIA systems might perform differently. Most DR screening programmes in well-resourced settings use staged mydriatic 2-field

imaging and human grading protocols, based on achieving diagnostic test accuracies of at least 80% sensitivity and 95% specificity (4). Whether this ARIA software may already have some utility in regular DR screening for people with diabetes in less-advantaged and under-resourced areas where access to and uptake of screening is limited is of interest, but not yet fully resolved. Retinal imaging and grading by trained retinal graders are widely accepted for DR screening (24), however lack of trained retinal graders physically in remote / very remote Australia is a major problem. Having to send images off-site for grading may delay reporting of DR status, potentially delaying treatment for those who need it, particularly as referrals to ophthalmology may require individuals travelling by plane or waiting for outreach ophthalmology services. For this reason and the high prevalence of DR in these populations, artificial intelligence for on-site primary image analysis during the DR screening process could play an important role.

DR prevalence in this indigenous primary-care setting was high, 47% (21) using the NHMRC clinical guidelines for DR management, based on the ETDRS classification for DR (2008) (25). In the current study, DR grading was by ARIA software (designed to grade to UK criteria), not yet approved for clinical use, although it has been trialled on other populations using different retinal fundus cameras (22), and a UK human grader using the UK national screening programme DR classification. At individual level, we report high sensitivity for ARIA in detecting any DR (91.4%). This is similar to that reported by Rajalakshmi *et al* (2018) where they used an alternate AI system, (EyeArt), at a tertiary care diabetes centre in India, capturing three retinal images using a smartphone-based device and reported a sensitivity of 95.8% for detection of any DR. Another study, using the Medios AI system, captured two retinal images and reported similar sensitivity and specificity values (26). The high sensitivity produced for any DR in this study suggests that the ARIA used is a useful primary-care tool for triaging to human graders and ultimately reducing the workload, time, cost and variability of human DR grading. Other studies used more retinal fields in their imaging protocol compared to the single-image protocol in our grading study. Despite this, we found a higher sensitivity in the detection of any DR. If only one retinal image proves sufficient for detecting any DR, this may further reduce cost and time to image and triage a patient to human grading. However, future studies are needed in which the imaging protocol remains the same i.e. a single image as herein vs. a 2-image protocol (e.g. as per the NHS DES Programme protocol) to test whether any DR is missed with a single image.

The main function of current grading systems is to identify referable cases requiring eye treatment and, increasingly, to identify people at risk of progression to ensure they are screened in a timely manner. In the current study the ARIA software had a sensitivity of 96.8% for DR grade R3, however this ARIA version didn't take into account a combined score for 'any referable DR', $\geq R2$, nor does it have separate maculopathy grading. Optical coherence tomography, the gold standard for diagnosing maculopathy, would ideally be used alongside ARIA, however, in remote Australia only portable OCT's would be practical, and these are not yet widely available in this region. Olson *et al* (2013) reported that OCT imaging in patients at higher risk of maculopathy prior to clinical examination, reduced health service costs, further supporting OCT integration into DR screening programmes (27).

ARIAs offer the opportunity to broaden the delivery of a much needed screening service, while also freeing resources for other priority areas in health-care, as well as potentially facilitating the organisation and facilitation of a DR screening service into less resourced settings. In this study, ARIA was used offsite by its developers in collaboration with the TEAMSnet study group. Should ARIA be used in geographically remote and very remote settings, it is preferable for the software to be on local computers, rather than 'in the cloud' as used in many studies (28-30), as internet connectivity is often episodic and less reliable in rural and remote regions. Furthermore, some cultural groups restrict health data going off-site.

More detailed evaluation is required for accurate interpretation of false positives and false negatives. Reasons for over-grading by ARIA were primarily the presence of other pathology and poor image quality, mainly due to cataracts and/or avoidable imaging artefacts. Another reason for 'over-grades' may be due to the presence of maculopathy, which is graded separately from retinopathy in the UK grading system. An advantage of arbitrating discordant images is that a human grader can discern reasons why ARIA over- or under-graded the retinopathy level, thereby providing important information for future software development. Of the false negatives found herein, none had sight-threatening DR.

Both study strengths and limitations exist: Unlike previous ARIA studies that used excellent image quality, this study analysed images of widely varying quality taken by locally trained primary-care health workers, thus is more representative of real-world primary-care DR screening. The study was a clinical study of a high DR risk group. An experienced certified human grader [the comparator] graded the images as per the UK national DR screening protocol. In addition, the original DR grading results of certified Australian graders were

made available for comparison with the UK human and ARIA grades: the frequency of DR and its various stages by Australian NHMRC grading criteria were comparable. Adjudication by two experienced retinal graders of ARIA false positives and negatives provides insights that may aid in future software development. The current ARIA software version may be suitable for desktop or local server use for triaging images to human grading. Study limitations include the lack of data on diabetic maculopathy, which is a very important sight-threatening complication of DR. Missing information on referable DR is another important limitation for the current version of this ARIA software. Only one (comparator) human grader was used in this study. Another limitation includes the diagnostics and evaluation of false positives and negatives requiring future studies.

Conclusion

This study presents a current ARIA software as a potentially useful tool in DR screening in remote and very remote Indigenous Australians for triaging screened participants into two groups – those without reduced vision and/or DR who should be screened locally according to national guidelines and the remainder with DR and/or impaired vision who require human grading for severity, referral and follow-up. This would reduce the workload of human graders by half in this clinical setting and population by eliminating those without DR or reduced vision. In due course, updated ARIA versions of this application may achieve adequate sensitivity for the detection of referable retinopathy and maculopathy. Human graders could then prioritise their efforts to the quality assurance of all elements of an ARIA-based DR screening program.

References

1. Lee R, Wong TY, Sabanayagam C. Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss. *Eye and vision*. 2015;2(1):17.
2. Association AD. Reduction in weight and cardiovascular disease risk factors in individuals with type 2 diabetes: one-year results of the look AHEAD trial. *Diabetes Care*. 2007;30(6):1374-83.
3. Thomas R, Halim S, Gurudas S, Sivaprasad S, Owens D. IDF Diabetes Atlas: A review of studies utilising retinal photography on the global prevalence of diabetes related retinopathy between 2015 and 2018. *Diabetes Res Clin Pract*. 2019;157:107840.
4. Scanlon PH. The English national screening programme for diabetic retinopathy 2003–2016.

Acta diabetologica. 2017;54(6):515-25.

5. Avidor D, Loewenstein A, Waisbourd M, Nutman A. Cost-effectiveness of diabetic retinopathy screening programs using telemedicine: a systematic review. *Cost Effectiveness and Resource Allocation*. 2020;18:1-9.
6. Bhaskaranand M, Ramachandra C, Bhat S, Cuadros J, Nittala MG, Sadda SR, et al. The value of automated diabetic retinopathy screening with the EyeArt system: a study of more than 100,000 consecutive encounters from people with diabetes. *Diabetes Technol Ther*. 2019;21(11):635-43.
7. Rajalakshmi R, Subashini R, Anjana RM, Mohan V. Automated diabetic retinopathy detection in smartphone-based fundus photography using artificial intelligence. *Eye*. 2018;32(6):1138.
8. Scotland GS, McNamee P, Fleming AD, Goatman KA, Philip S, Prescott GJ, et al. Costs and consequences of automated algorithms versus manual grading for the detection of referable diabetic retinopathy. *Br J Ophthalmol*. 2010;94(6):712-9.
9. Bouhaimed M, Gibbins R, Owens D. Automated detection of diabetic retinopathy: results of a screening study. *Diabetes Technol Ther*. 2008;10(2):142-8.
10. Jelinek HJ, Cree MJ, Worsley D, Luckie A, Nixon P. An automated microaneurysm detector as a tool for identification of diabetic retinopathy in rural optometric practice. *Clinical and Experimental Optometry*. 2006;89(5):299-305.
11. Niemeijer M, van Ginneken B, Russell SR, Suttorp-Schulten MS, Abramoff MD. Automated detection and differentiation of drusen, exudates, and cotton-wool spots in digital color fundus photographs for diabetic retinopathy diagnosis. *Invest Ophthalmol Vis Sci*. 2007;48(5):2260-7.
12. Philip S, Fleming AD, Goatman KA, Fonseca S, Mcnamee P, Scotland GS, et al. The efficacy of automated “disease/no disease” grading for diabetic retinopathy in a systematic screening programme. *Br J Ophthalmol*. 2007;91(11):1512-7.
13. Welikala R, Dehmeshki J, Hoppe A, Tah V, Mann S, Williamson TH, et al. Automated detection of proliferative diabetic retinopathy using a modified line operator and dual classification. *Comput Methods Programs Biomed*. 2014;114(3):247-61.
14. Fleming AD, Goatman KA, Philip S, Williams GJ, Prescott GJ, Scotland GS, et al. The role of haemorrhage and exudate detection in automated grading of diabetic retinopathy. *Br J Ophthalmol*. 2010;94(6):706-11.
15. Tufail A, Rudisill C, Egan C, Kapetanakis VV, Salas-Vega S, Owen CG, et al. Automated diabetic retinopathy image assessment software: diagnostic accuracy and cost-effectiveness compared with human graders. *Ophthalmology*. 2017;124(3):343-51.
16. Fleming AD, Goatman KA, Philip S, Prescott GJ, Sharp PF, Olson JA. Automated grading for diabetic retinopathy: a large-scale audit using arbitration by clinical experts. *Br J Ophthalmol*. 2010;94(12):1606-10.
17. Tufail A, Kapetanakis VV, Salas-Vega S, Egan C, Rudisill C, Owen CG, et al. An observational study to assess if automated diabetic retinopathy image assessment software can replace

- one or more steps of manual imaging grading and to determine their cost-effectiveness. *Health Technol Assess.* 2016;20(92):1-72.
18. Brazionis L, Jenkins A, Keech A, Ryan C, Bursell S-E. An evaluation of the telehealth facilitation of diabetes and cardiovascular care in remote Australian Indigenous communities:- protocol for the telehealth eye and associated medical services network [TEAMSnet] project, a pre-post study design. *BMC Health Serv Res.* 2017;17(1):1-12.
19. Statistics ABo. Australian Standard Geographical Classification (ASGC) 2006 [updated 27/09/2007. Available from: <https://www.abs.gov.au/AUSSTATS/abs@.nsf/Latestproducts/1AEE5890BD5B7B5ECA2571A90017901F?opendocument>.
20. Bourne RR, Flaxman SR, Braithwaite T, Cicinelli MV, Das A, Jonas JB, et al. Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis. *The Lancet Global Health.* 2017;5(9):e888-e97.
21. Brazionis L, Jenkins A, Keech A, Ryan C, Brown A, Boffa J, et al. Diabetic retinopathy in a remote Indigenous primary healthcare population: a Central Australian diabetic retinopathy screening study in the Telehealth Eye and Associated Medical Services Network project. *Diabetic Medicine.* 2018;35(5):630-9.
22. Wakerman J, Humphreys J, Russell D, Guthridge S, Bourke L, Dunbar T, et al. Remote health workforce turnover and retention: what are the policy and practice priorities? *Human resources for health.* 2019;17(1):1-8.
23. Al Turk L, Wang S, Krause P, Wawrzynski J, Saleh GM, Alsawadi H, et al. Evidence based prediction and progression monitoring on retinal images from three nations. *Translational Vision Science & Technology.* 2020;9(2):44-.
24. Singer DE, Nathan DM, Fogel HA, Schachat AP. Screening for diabetic retinopathy. *Annals of Internal Medicine.* 1992;116(8):660-71.
25. Mitchell P, Foran S, Wong T, Chua B, Patel I, Ojaimi E, et al. Guidelines for the management of diabetic retinopathy. National Health and Medical Research Council. 2008.
26. Sosale B, Aravind SR, Murthy H, Narayana S, Sharma U, Gowda SG, et al. Simple, Mobile-based Artificial Intelligence Algorithm in the detection of Diabetic Retinopathy (SMART) study. *BMJ Open Diabetes Research and Care.* 2020;8(1).
27. Olson J, Sharp P, Goatman K, Prescott G, Scotland G, Fleming A, et al. Improving the economic value of photographic screening for optical coherence tomography-detectable macular oedema: a prospective, multicentre, UK study. *Health Technology Assessment.* 2013.
28. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama.* 2016;316(22):2402-10.
29. Ting DSW, Cheung CY-L, Lim G, Tan GSW, Quang ND, Gan A, 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(22):2211-23.
30. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based
diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ digital medicine*.
2018;1(1):1-8.

Author Manuscript

Table 1 ARIA diagnostic performance relative to a human grading comparator* for diabetic retinopathy classification.

A. By participant

ARIA	Human grading				Ungradable n (%)
	R0 n (%)	R1 n (%)	R2 n (%)	R3 n (%)	
R0 n (%)	175 (85.0)	26 (12.6)	4 (1.9)	1 (0.5)	0 (0)
R1 n (%)	16 (13.8)	82 (70.7)	18 (15.5)	0 (0)	0 (0)
R2 n (%)	0 (0)	2 (5.3)	31 (81.6)	5 (13.2)	0 (0)
R3 n (%)	0 (0)	0 (0)	1 (3.2)	30 (96.8)	0 (0)
Ungradable n (%)	0 (0)	0 (0)	1 (6.7)	1 (6.7)	13 (86.7)

B. Sensitivity and specificity by eye and participant.

Comparator DR	All available eyes	Participant worse eye
---------------	--------------------	-----------------------

status	ARIA Sensitivity % (95%CI)	ARIA Specificity % (95%CI)	ARIA Sensitivity % (95%CI)	ARIA Specificity % (95%CI)
Any DR	91.7 (88.0 - 94.6)	90.5 (87.3 - 93.0)	91.4 (86.3 – 95.0)	85.0 (79.3 - 89.5)
R1	74.6 (67.8 - 80.6)	93.3 (90.9 - 95.3)	70.7 (61.5 - 78.8)	89.8 (85.6 - 93.1)
R2	83.6 (72.5 - 91.5)	95.4 (93.5 - 96.9)	81.6 (65.7 - 92.3)	93.5 (90.4 - 95.8)
R3	95.7 (85.1 - 99.5)	98.6 (97.3 - 99.3)	96.8 (83.3 - 99.9)	98.3 (96.4 - 99.4)

Abbreviations: ARIA

Definitions: R0: no apparent DR (ETDRS Level 10; DR absent), R1: background DR (ETDRS Mild NPDR Level 20 & Moderate NPDR Level 35-43), R2: pre-proliferative DR (ETDRS Moderate NPDR Level 47 & Severe NPDR Level 53), R3: proliferative DR (ETDRS PDR Level 61-85)

* *Reference standard* ^ Percentages are within overall human grading

Table 2 Number of participants with both eyes the same DR grade, only one eye the same grade, both eyes different grade and DR grades of participants with only one eye available for analysis for ARIA and human comparator

	Human grade	ARIA grade	Number
Number of participants with both eyes the same grade			317
Only 1 eye the same grade			57
Second eye worse in ARIA than in human			42
	<i>R0</i>	<i>R1</i>	19
		<i>R2</i>	3
		<i>R3</i>	0
		<i>UG</i>	0
	<i>R1</i>	<i>R2</i>	10
		<i>R3</i>	0
		<i>UG</i>	0
	<i>R2</i>	<i>R3</i>	3
		<i>UG</i>	0
Human grade UG, ARIA grade R2			7
Second eye better in ARIA than in human			15

	<i>R1</i>	<i>R0</i>	11
	<i>R2</i>	<i>R1</i>	2
	<i>R3</i>	<i>R2</i>	2
Both eyes different grade			26
Both eyes worse in ARIA than in human			19
	<i>R0R0</i>	<i>R2R2</i>	1
	<i>R0R0</i>	<i>R3R2</i>	1
	<i>R0R0</i>	<i>R1R1</i>	8
	<i>R1R1</i>	<i>R2R2</i>	6
	<i>R2R2</i>	<i>R3R3</i>	3
Both eyes better in ARIA than in human			7
	<i>R1R1</i>	<i>R0R0</i>	7
10 subjects with only a single eye available for analysis			10
Eyes were the same in ARIA and human			8
	<i>R0</i>	<i>R0</i>	5
	<i>R1</i>	<i>R1</i>	2
	<i>R2</i>	<i>R2</i>	1
Eyes ARIA was worse than human	<i>R1</i>	<i>R2</i>	1
Eyes human was UG whereas ARIA graded it as R3	<i>UG</i>	<i>R3</i>	1
		Total	410

Abbreviations: ARIA (Automated Retinal Image Analysis), UG (Ungradable)

Definitions: Definitions: R0: no apparent DR (ETDRS Level 10; DR absent), R1: background DR (ETDRS Mild NPDR Level 20 & Moderate NPDR Level 35-43), R2: pre-proliferative DR (ETDRS Moderate NPDR Level 47 & Severe NPDR Level 53), R3: proliferative DR (ETDRS PDR Level 61-85)

Author Manuscript