Meredith Sarah (Orcid ID: 0000-0003-3563-468X)

Corresponding author mail id: sarah.meredith@health-intelligence.com

Performance of an artificial intelligence automated system for diabetic eye screening in a large English population

Sarah Meredith[1], Mark van Grinsven[2], Jonne Engelberts[2], Dominic Clarke[1], Vicki Prior[1], Jo Vodrey[1], Alison Hammond[1], Raja Muhammed[1], Philip Kirby[1]

1. InHealth Intelligence Ltd, UK

2. Thirona B.V. The Netherlands

Corresponding author

Sarah Meredith, ORCID ID: 0000-0003-3563-468X

Abstract

Aims

A diabetic eye screening programme has huge value in reducing avoidable sight loss by identifying diabetic retinopathy at a stage when it can be treated. Artificial intelligence automated systems can be used for diabetic eye screening but are not employed in the national English Diabetic Eye Screening Programme. The aim was to report the performance of a commercially available deep learning artificial intelligence software in a large English population.

Methods

9,817 anonymised image sets from 10,000 consecutive diabetic eye screening episodes were presented to an artificial intelligence software. The sensitivity and specificity of the artificial intelligence system for detecting diabetic retinopathy was determined using the diabetic eye screening programme manual grade according to national protocols as the reference standard.

Results

For no diabetic retinopathy vs any diabetic retinopathy the sensitivity of the artificial intelligence grading system was 69.7% and specificity 92.2%. The performance of the artificial intelligence system was superior for no or mild diabetic retinopathy vs significant or referrable diabetic retinopathy with a sensitivity of 95.4% and specificity of 92.0%. No cases were identified in which the artificial intelligence grade had missed significant diabetic retinopathy.

Conclusion

The performance of a commercially available deep learning artificial intelligence system for identifying diabetic retinopathy in an English national Diabetic Eye Screening Programme is

This article has been accepted for publication and 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.15055

presented. Using the pre-defined settings artificial intelligence performance was highest when identifying diabetic retinopathy which requires an action by the screening programme.

Novelty statement

What is already known:

Machine learning artificial intelligence systems can be used effectively for diabetic eye screening.

What this study has found:

The performance of a commercially available deep learning artificial intelligence system was 69.7% sensitive and 92.2% specific in an English population for detecting any diabetic retinopathy present, and more sensitive at 95.4% with specificity 92.0% for distinguishing between mild and significant retinopathy which requires a potential action by the screening programme.

What are the implications of the study:

Adds to the evidence base that a deep learning artificial intelligence system could be a safe resource in the context of English diabetic eye screening, and there is value in further investigating where an automated system may be best employed.

Introduction

Diabetic retinopathy is a microvascular complication of diabetes and eye screening using digital retinal photography is one of the nine key processes for diabetic care in the UK.[1,2] England has a well-established national Diabetic Eye Screening Programme (DESP) whose key aim is to reduce sight loss in people with diabetes by the prompt identification of sight-threatening retinopathy at an appropriate stage.

The identification of diabetic retinopathy (DR) has already undergone a major shift from hospital-based evaluation by an ophthalmologist to a modern telemedicine process of asynchronous store-and-forward using retinal photography. Potential retinal photograph image analysis approaches are also progressing but haven't been adopted yet by English screening programmes. Diabetic eye screening is labour intensive.[3] An image set which results in a referral to the hospital eye service is typically independently graded by three trained graders before the outcome is determined. Artificial intelligence (AI) in healthcare has been applied to a number of medical imaging settings including analysis of the modern standard retinal photographs used by the English DESP.[4]

The Scottish DESP has been using an automated system for the first grading of retinal images since 2011[5] as has the Portuguese DESP.[6] Following a study on 900 people with diabetes a deep learning system had US Food and Drug Administration approval for the diagnosis of DR.[7] The AI based software currently being used in DR screening programmes are machine learning systems which recognise retinal microaneurysms and separate images with no DR from those with any DR. The incorporation of deep learning algorithms in more recently developed AI based systems is likely to increase accuracy[8], and also introduces the potential to differentiate between retinopathy which can be monitored by the screening service and referrable retinopathy which requires further hospital investigation for treatment.

The English DESP has extensive quality assurance checks built into the grading process. All human graders independently grading are required to consistently demonstrate a sensitivity of over 85% and a specificity of over 80% for identifying referrable DR.[9] This is measured using standardised monthly test and training images as well as regular assessment of peer agreement within the programme.

It is necessary for a technology to demonstrate evidence of the accuracy of an automated system for both the clinical setting and relevant population in which it's proposing it may be employed.[10] The aim of this study is to report the performance of a commercially available deep learning based software for the detection of DR in an English screening programme by comparing the automated grading outcomes to the manual grading DESP classification, and to provide an opportunity to verify the algorithm was optimised for this population.

Methods

In April and May 2021 90.1% of people attending a North West London DESP appointment provided consent for their anonymised images to be used for the purposes of research. Anonymisation consisted specifically of removing all personal data and all details of the screening visit. The DESP outcome was preserved as the reference standard. Inclusion of images in the dataset presented to the automated grading system did not change any clinical pathways, which had already been determined by the DESP manual grade. The Research Collaboration Agreement included confidentiality and data protection clauses. A comprehensive Data Protection Impact Assessment (DPIA) was prospectively conducted before any images were released to the automatic grading system. The Standards for Reporting of Diagnostic Accuracy (STARD) statement items have been reviewed (checklist provided as supplementary material).[11]

The demographics of the programme population for 2021 – 2022 and the demographics of the study population over these two months are presented in table 1 following the NHS Digital guidance on recording ethnicity.[12]

Table 1

Activity and uptake of the North West London (NWL) Diabetic Eye Screening Programme 2021 – 2022 with the demographics of those who attended a routine digital screening (RDS) appointment and the demographics of the study population.

	NWL screening population (2021/22)	Percentage	Study population	Percentage
Patients invited (RDS)	137,909			
Screened (RDS)	105,702	76.6%	10,000	
Ethnicity				
White	28,068	26.6%	2,672	26.7%
Asian	50,632	47.9%	4,958	49.6%
Black	11,757	11.1%	1,017	10.2%
Mixed	2,280	2.2%	209	2.1%
Other	7,582	7.2%	652	6.5%
Not-stated	5,383	5.1%	492	4.9%
Gender (only male/female included)				
Male	57,102	54.0%	5,344	53.4%
Female	48,597	46.0%	4,656	46.6%
Age				
Mean	61 (range 12 to 108)		62 (range 12 to 100)	

People who were ungradable in one and both eyes have been excluded from the reported analysis. 9,817 image sets from 10,000 consecutive, and consented, screening appointments over two months were presented to the automated system. The automated AI system gave a result for every image set presented to it.

The software used for the automated detection of DR was RetCAD v.2.1.0 (Thirona, The Netherlands). This is a commercially available Class IIa CE-marked medical device software that incorporates a deep learning framework for the detection of abnormalities in colour fundus images.[13] The software is based on convolutional neural networks for the tasks of recognising

diabetic retinopathy, age-related macular degeneration, and glaucoma. As the focus of this analysis was on the detection of DR only the DR component of the software was used.

The DR component of the software consists of an ensemble of convolutional neural networks predicting a severity score for DR in a fundus image. Each of the individual deep learning networks in the ensemble were trained end-to-end by providing an input fundus image and target reference severity label according to the International Clinical Diabetic Retinopathy Disease Severity Scale (ICDR).[14] None of the images used in this study were used to train or develop the AI system. After training the individual networks each produce a regression score for the image on a continuous scale from 0 to 4 according to the ICDR classification and the average of the individual predictions is the final outcome DR score of the system.

The DESP final manual grade is the reference standard used to compare. DESP require a macular retinal image and nasal image to be taken for each eye at the screening appointment resulting in four images from each person being presented to a manual grader. The English DESP grading process is features based to determine the DESP grade.[15] The automated system is compared only to the retinopathy 'R' grade of the DESP features based grading classification. The maculopathy 'M' grade was not taken into account. Table 2 maps the RetCAD AI score to the ICDR level and DESP grade.

Table 2

Mapping of RetCAD v.2.1.0 Artificial Intelligence (AI) score to the International Clinical Diabetic Retinopathy Disease Severity Scale (ICDR) and Diabetic Eye Screening Programme (DESP) grade.

RetCAD AI score	ICDR level	DESP grade
Stage 0: 0-0.5	Stage 0: No DR	R0: No retinopathy
Stage 1: 0.5-1.5	Stage 1: Mild DR	R1: Background retinopathy
Stage 2: 1.5-2.5	Stage 2: Moderate DR	R2: Pre-proliferative retinopathy
Stage 3: 2.5-3.5	Stage 3: Severe DR	R2: Pre-proliferative retinopathy
Stage 4: 3.5-4.0	Stage 4: Proliferative DR	R3(a/s): Proliferative retinopathy

Images with a DESP grade of R2 and R3 (either active 'a' or stable 's') are deemed to have significant retinopathy and, within DESP, will have a further referral outcome grade by an experienced grader which often results in a referral to the hospital eye service. The North West London DESP grading and referral process is managed within the inhouse developed software Spectra.

The results presented in this study are at a person level (table 3) with eye-level data agreement provided as a supplemental table. To generate the person level grade the most severe manual reference grade (determined using the four images obtained at the screening encounter) was taken as the person level grade. For the AI system a similar approach was used by determining the worst DR score (highest DR severity score) as the person level DR score. As the automated analysis was run in parallel with, and didn't alter, the normal screening pathway all those with a final manual DESP grade of R2 and R3 had either a hospital referral or were transferred onto a frequent monitoring pathway following national guidance.[15] Both a hospital referral and a transfer onto a different pathway are referred to as a pathway change in this paper.

The receiver operating characteristic (ROC) curve was created by plotting the sensitivity (true positive rate) against the specificity (false positive rate) at various pre-defined threshold settings of the severity score of the AI. Retrospective analysis then enables an optimal threshold to be determined which is the cut-off point with the best trade-off between sensitivity and specificity (point closest to the upper left corner of the ROC curve). The results at the pre-defined cut-off are presented in this paper as this is what would be obtained should the software be deployed. The optimal threshold in this population is also reported for comparison. The 95% confidence interval of the curve was computed using bootstrap analysis. The area under a ROC curve (AUC) is an indication of the accuracy of a diagnostic test with a value of over 0.8 deemed good.[16]

In order to examine the impact disagreements between the manual and automated grade might have on the outcome of the screening visits image sets with a significant disagreement were retrospectively reviewed by a senior, experienced grader who doesn't work within the NorthWest London DESP and had not previously seen the images. A significant disagreement was considered to be one in which either the manual grade or automated grade would result in a pathway change.

Results

Of 10,000 screening appointments 183 were ungradable by a human grader resulting in 9,817 screening encounters presented to the automated AI system for analysis. The demographics of the study population and general population are very similar (table 1). The prevalence of any DR in this dataset was 27.11% with the total in each DESP grade retinopathy category presented in table 3.

Table 3

Patient level comparison of the manual diabetic eye screening programme (DESP) grade and the RetCAD artificial intelligence grade for diabetic retinopathy (DR) using the pre-defined cut-offs, with total number of image sets and percentage of image sets analysed for each DESP grade.

Manual DESP grade in worst affected eye	No DR	Mild DR	Moderate DR	Severe DR	Proliferative DR	Total number of image sets (percentage)
R0	6,600	495	54	7	0	7,156 (72.89%)
R1 (one eye)	691	632	205	4	1	1,533 (15.62%)
R1 (both eyes)	115	414	504	5	3	1,041 (10.60%)
R2	1	3	39	29	0	72 (0.73%)
R3 active	0	0	0	5	1	6 (0.06%)
R3 stable	0	0	1	2	6	9 (0.09%)

Figure 1 shows the ROC curves for no DR vs any DR (equivalent to DESP manual grade R0 vs R1/R2/R3) on the lower blue line and non-referrable DR vs referrable DR on the higher orange line. The orange curve therefore represents grades which had an outcome of annual recall in screening vs grades which required a pathway change (R0/R1 vs R2/R3). The shaded areas on the ROC curves are the 95% confidence intervals. The spread of the automated system DR severity scores by reference DESP grade is shown as a box plot in figure 2.

Figure 1

Receiver operating characteristic (ROC) curve of the RetCAD v.2.1.0 system for no diabetic retinopathy (DR) vs any DR (blue line) and no or mild DR vs referrable DR (orange line).

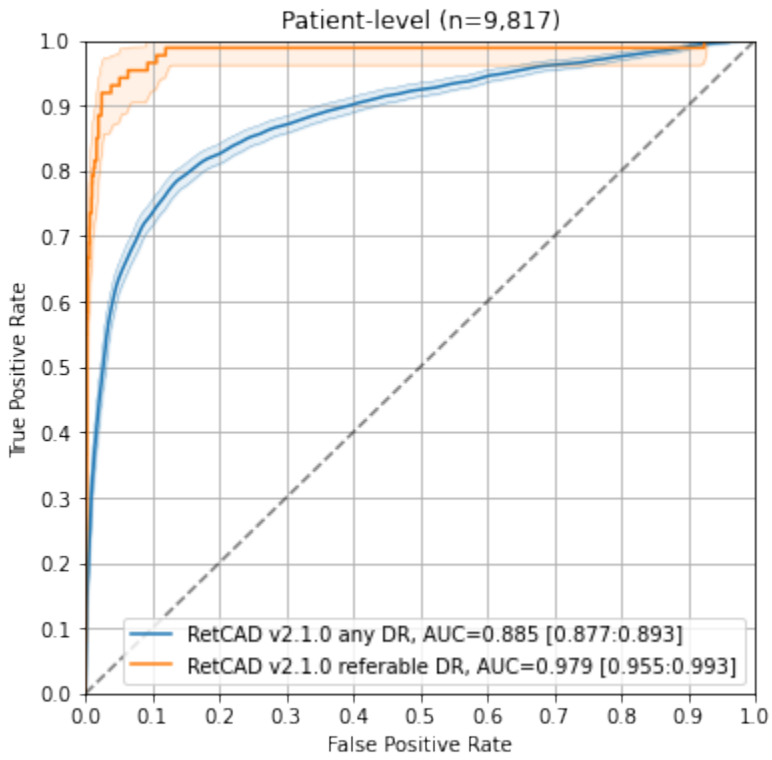
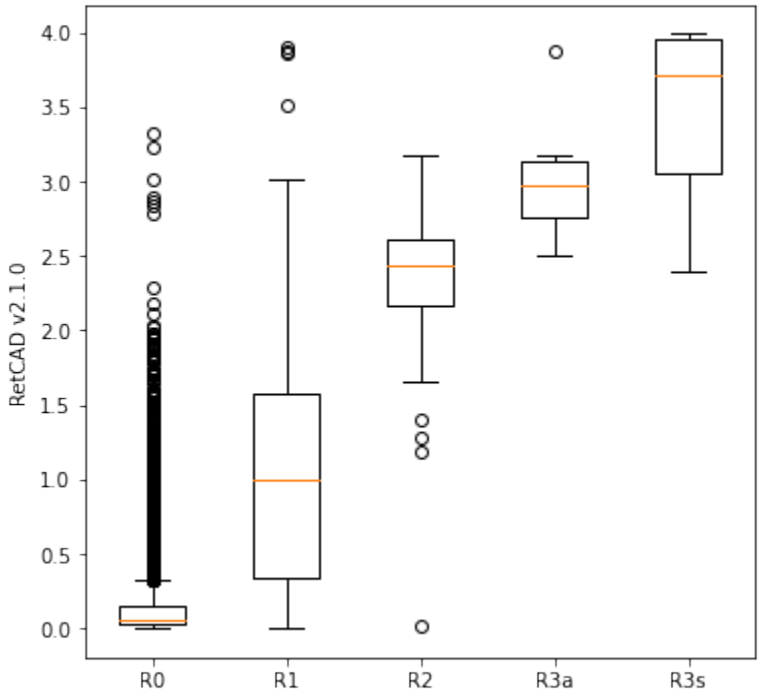


Figure 2

Box plot of the RetCAD v.2.1.0 diabetic retinopathy (DR) severity scores per diabetic eye screening programme (DESP) grade.



The comparison of the manual grade in the worst affected eye with the automated grading result is presented in table 3 at a person level. Eye level data is provided as a supplemental table. Using the predefined severity score of under 0.50 as the threshold the sensitivity of the AI based grading software to detect any DR is 69.7% with a specificity of 92.2%. If RetCAD identified a patient as having diabetic retinopathy the AI was correct in 76.9% indicating a precision (positive predictive value) of 76.9% (1854 cases out of 2410). At the optimal threshold of 0.30 for this population when detecting any DR the sensitivity is 78.5% and specificity 86.5%. In the no DR vs any DR analysis the AUC is 0.885 (95% confidence interval 0.877-0.893).

When differentiating between non-referrable DR (outcome annual recall by the screening service) vs referrable DR (outcome potentially a pathway change) the performance of the AI system at the predefined cut-off of 1.50 was 95.4% sensitivity and specificity 92.0%. The precision is 9.6% (83 positive cases out of 866). At the optimal threshold of 2.01 the sensitivity is 92.0% and specificity 97.6%. For non-referrable vs referrable DR the AUC is 0.979 (95% confidence interval 0.955 – 0.993) representing excellent diagnostic accuracy of the AI based system for detecting DR which potentially required an action by the screening programme.

For each ethnic group the number of people ungradable on the manual DESP grade and therefore excluded from the study population is presented in table 4 along with the number of people in whom the RetCAD AI disagreed with the manual grade.

Table 4

Number of people attending routine digital screening (RDS) who were ungradable on the reference Diabetic Eye Screening Programme (DESP) manual grade, and the number of disagreements of the RetCAD Artificial Intelligence (AI) outcome by ethnic group.

	Study population (Total in each ethnicity)	Ungradable (by manual reference DESP grade)	No DR vs any DR disagreement	Non-referrable DR vs referrable DR disagreement
White	2,672	45 (1.68%)	333 (12.46%)	191 (7.15%)
Mixed	209	8 (3.83%)	27 (12.92%)	13 (6.22%)
Asian	4,958	80 (1.61%)	702 (14.16%)	395 (7.97%)
Black	1,017	22 (2.16%)	140 (13.77%)	83 (8.16%)
Other	652	16 (2.45%)	103 (15.80%)	64 (9.82%)
Not stated	492	12 (2.44%)	58 (11.79%)	41 (8.33%)
Total	10,000	183 (1.83%)	1363 (13.88%)	787 (8.02%)

There were 24 significant disagreements between the reference DESP outcome grade and AI RetCAD grade. 20 represented an overgrade by the AI, and 4 an apparent undergrade in which the screening programme made a pathway change but the AI grade would have recalled to screening in one year. The details of the significant disagreements are presented in table 5. Of the 4 cases in which the AI apparently undergraded a second DESP regrade agreed with the AI outcome. Of the 24 disagreements manually regraded 15 had retinal signs of ocular pathology not related to DR.

Table 5

Summary of the disagreements between the reference Diabetic Eye Screening Programme (DESP) grade outcome and artificial intelligence (AI) RetCAD v.2.1.0 severity score generated outcome, with retrospective DESP regrade and comments.

	Reference standard DESP grade	RetCAD severity score	DESP regrade	Comment	Over / under grade by AI	Over / under grade by DESP
1	R0	3.2	R0	Camera artefact present	Over	
2	R1	2.6	R1		Over	
3	R1	2.5	R1	Referral made for macular DR	Over	
4	R1	3.8	R3 stable			Under
5	R1	2.6	R1	Non-DR signs	Over	
6	R0	3.0	R0	Non-DR signs	Over	
7	R1	3.8	R1	Non-DR signs	Over	
8	R1	2.8	R1	Non-DR signs	Over	
9	R1	2.8	R1	Non-DR signs	Over	
10	R1	3.9	R1	Non-DR signs	Over	
11	R1	3.0	R1	Non-DR signs	Over	
12	R0	2.7	R0	Non-DR signs	Over	
13	R0	2.8	R0	Non-DR signs	Over	
14	R0	2.6	R0	Non-DR signs	Over	
15	R1	2.6	R1		Over	
16	R0	3.3	R0		Over	
17	R1	3.5	R1	Non-DR signs	Over	
18	R1	2.6	R1	Pale atrophic retina	Over	
19	R1	2.9	R1	Non-DR signs	Over	
20	R0	2.8	R0	Non-DR signs	Over	
21	R2	0.0	R0			Over
22	R2	1.3	R1	Non-DR signs		Over
23	R2	1.1	R1	Non-DR signs		Over

24	R2	1.2	R1			Over

Discussion

This paper presents the performance of a deep learning AI system for diabetic eye screening in a large, representative, real-world screening population in North West London. The proportion of images that had moderate retinopathy (R2) and proliferative retinopathy (R3) in this cohort were 0.7% and 0.2% respectively. This is lower than has been reported previously in England[3] but is the real-world scenario of two months of activity for this screening programme. Analysing a specific sample of severe DR (R3) using more months of screening would provide a greater evidence base and is an option for future research.

As well as presenting the performance of the RetCAD software to detect referrable patients we also present results for no DR versus any DR. The difference between these protocols is the inclusion of the mild DR category as a positive test outcome. This would enable patients to be detected at an early stage of retinopathy. Different pathways for patients with no DR and mild DR are possible in the form of a more frequent screening interval.

There were disagreements between the automated and manual grade. This is not unexpected. It is accepted a screening programme will not deliver 100% sensitivity and missed pathology is expected from any grading system. In the data presented the amount of disagreement between the automated AI and manual grade appeared to be similar in different ethnicity groups, but this has not been investigated in detail and is an area for future research. In the patient level comparison of manual and AI grade there appears to be a difference between R1 in one eye and R1 in both eyes. This could be interpreted as the AI appearing more likely to grade as moderate DR if R1 is present in both eyes while the manual grader using a features based protocol was still classifying as mild DR. Most of the disagreements that were retrospectively reviewed as part of this paper were overgrading by the AI. This is reassuring but it's recognised if only AI were used in real-world practice the performance results do not state how many people, with hindsight, clinically required referring. The reference standard of normal best practice also doesn't correctly identify all cases.

Using the pre-defined cut-offs the RetCAD AI system achieved a sensitivity of 69.7% with a specificity of 92.2% for detecting any DR. A manual grader would be identified as someone who required retraining if this sensitivity was observed. The advantage of having a continuous score for the AI output is the desired sensitivity of the AI can be altered by changing the cut-off point. Choosing an increased sensitivity will result in a decreased specificity. It's at a first level grading stage of screening the older machine learning systems are currently employed and it's thought at present the most cost-effective implementation of an automated grading system would be at this no retinopathy vs any retinopathy level.[17] The cut-off point of the software can be optimised for a population and set to obtain a workload reduction of 50% (i.e. a specificity of 50% at a sensitivity of 92.4%) as a level for the detection of any DR which is reported to be cost-effective.[17]

Whereas DESP has experience of, and is comfortable with, retraining human graders how a programme would 'retrain' an automated system beyond identifying a cut-off point of an AI system which is optimised for the relevant population is not clear. The manual reference DESP grade is a features-based classification. In general, end-to-end trained deep learning systems are 'black-box' systems. There is no direct relation of the output DR score to the presence of human characterised features. There is evidence black-box systems are triggered by presence of these features, but the system is not specifically trained to identify features. In an end-to-end system the target label is provided, and the network will work out which parts of the images are important to come to this decision. There remain potential challenges for DESP in addressing issues when misgrades by an automated system are identified.

The RetCAD deep learning system results we've presented show a superior performance at a no/mild DR vs significant DR level in this population. A sensitivity of 95.4% and specificity of 92.0% for non-referrable vs referrable DR far exceeds the grading quality assurance standards expected of a human grader raising the potential of a further place for AI in a screening programme as part of internal quality assurance processes.

A tendency for an overgrade to occur was seen in images with retinal signs of other ocular conditions not related to DR for both AI and the first manual grader. No cases were found in which the AI had significantly undergraded. If the AI system were used as a filter at a no DR vs any DR first level of grading an overgrade would not cause a pathway change as it would simply highlight to the second level human grader something is present. There would then be the same opportunity for the outcome to appropriately reflect what that something was.

In practice automated grading would be incorporated into the existing quality assurances of DESP which would allow programmes to measure sensitivity and specificity. This would contemporaneously highlight any differences and, more importantly, concerns regarding grading allowing the screening programme to address these.

A diabetic eye screening programme has huge value reducing avoidable sight loss which is well documented.[18] However, the results presented in this paper also highlight the number of completely normal retinal images that are seen by DESP. 14,510 eyes in two months of one regional screening programme in England were graded and no DR seen. 1,451 of these images were regraded for quality assurance of the first manual grade. The volume of grading required by DESP is such that training, experience, and maintaining skills for human graders is not predicted to be an issue if some grading were replaced by AI.

Irrespective of where AI could potentially be implemented into a diabetic eye screening programme – for quality assurance, first level filter grading, or identifying sight-threatening referrable DR - should AI based software be used in the diabetic eye screening pathway people attending a diabetic eye screening clinic must be fully informed and know that part of the analysis of their retinal images will be conducted by an automated system. The next step is to conduct a cost-benefit analysis of the RetCAD system at the different grading levels within DESP, with the aim of safely reducing the number of manual grades required and better utilising the specialist skills of our dedicated human graders.

Conclusion

AI systems are already being used for diabetic eye screening but not employed by English screening programmes. The performance of the RetCAD system for detecting DR is presented in an English population demonstrating a sensitivity of 69.7% and specificity 92.2% prior to optimisation. The deep learning algorithm demonstrated a superior performance when distinguishing between no or mild DR and significant retinopathy which requires the screening programme to make a pathway change with sensitivity 95.4% and specificity 92.0%. The evidence base examining deep learning AI systems as a safe resource in the context of diabetic eye screening is increasing.

References

1. Type 1 diabetes in adults: diagnosis and management NICE guideline (NG17) 2022. https://www.nice.org.uk/guidance/ng17. Accessed Sept 9, 2022.
2. Type 2 diabetes in adults: management NICE guideline (NG28) 2022. https://www.nice.org.uk/guidance/ng28. Accessed Sept 9, 2022.
3. Heydon P, Egan C, Bolter L, Chambers R, Anderson J, Aldington S, Stratton IM, Scanlon PH, Webster L, Mann S, du Chemin A, Owen CG, Tufail A, Rudnicka AR. Prospective evaluation of an artificial intelligence enabled algorithm for automated diabetic retinopathy screening of 30 000 patients. Br J Ophthalmol. 2021;105:723–728.
4. Bellemo V, Lim G, Rim TH, Tan GSW, Cheung CY, Sadda S, He M, Tufail A, Lee ML, Hsu W, Ting D. Artificial Intelligence Screening for Diabetic Retinopathy: the Real-World Emerging Application. *Curr Diab Rep*. 2019;19:72 .
5. Philip S, Lee N, Black M, Sharp P, Olson J. Impact of introducing automated grading into the Scottish national diabetic retinopathy screening programme. *Diabetic Medicine*. 2017;34 (Supplement 1):172.
6. Ribeiro L, Oliveira CM, Neves C, Ramos JD, Ferreira H, Cunha-Vaz J. Screening for Diabetic Retinopathy in the Central Region of Portugal. Added Value of Automated 'Disease/No Disease' Grading. *Ophthalmologica.* 2015;233:96-103 .
7. Abramoff 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;1:1-8.
8. Ting DSW, Pasquale LR, Peng L, Peng L, Campbell JP, Lee AY, Raman R, Tan GSW, Schmetterer L, Keane PA, Wong TY. Artificial intelligence and deep learning in ophthalmology. *Br J Ophthalmol.* 2019;103:167-175.
9. The management of grading quality: Good practice in the quality assurance of grading 2016. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/512832/The_Management_of_Grading.pdf. Accessed Nov 18, 2022.
10. UK National Screening Committee (UK NSC). Automated grading in the Diabetic Eye Screening Programme External review against programme appraisal criteria for the UK National Screening Committee 2021. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1035903/Evidence_summary_AI_in_DESP_2021.pdf. Accessed Sept 9, 2022.
11. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, Lijmer JG, Moher D, Rennie D, de Vet HCW, Kressel HY, Rifai N, Golub RM, Altman DG, Hooft L, Korevaar DA, Cohen JF, For the STARD Group. STARD 015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies. BMJ. 2015;351:h5527. PMID: 26511519d
12. Ethnicity 2022. https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-sets/mental-health-services-data-set/submit-data/data-quality-of-protected-characteristics-and-other-vulnerable-groups/ethnicity. Accessed Nov 18, 2022.

13. González-Gonzalo C, Sánchez-Gutiérrez V, Hernández-Martínez P, Contreras I, Lechanteur YT, Domanian A, van Ginneken B, Sánchez CI. Evaluation of a deep learning system for the joint automated detection of diabetic retinopathy and age-related macular degeneration. *Acta Ophthalmol*. 2020;98(4):368-377.
14. Wilkinson CP, Ferris FL 3rd, Klein RE, Lee PP, Agardh CD, Davis M, Dills D, Kampik A, Pararajasegaram R, Verdaguer JT; Global Diabetic Retinopathy Project Group. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*. 2003; 110: 1677–1682
15. NHS Diabetic Eye Screening Programme: grading definitions for referable disease 2021. https://www.gov.uk/government/publications/diabetic-eye-screening-retinal-image-grading-criteria/nhs-diabetic-eye-screening-programme-grading-definitions-for-referable-disease. Accessed Sept 9, 2022.
16. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982:143(1):29-36.
17. Tufail A, Kapetanakis VV, Salas-Vega S, 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;20:1–72.
18. Mohamed Q, Gillies MC, Wong TY. Management of diabetic retinopathy: a systematic review. *JAMA*. 2007;298:902–16.

Conflict of interests disclosure

SM, DC, VP, JV, AH and RM are all employees of InHealth Intelligence and PK is the Managing Director of InHealth Intelligence which is contracted to deliver Diabetic Eye Screening Programmes (for over 1.1 million people in England). JE is an employee of Thirona, and MvG is an employee and shareholder of Thirona.

Ethic statement

The study conforms to recognized standards. We've checked with the Health Research Authority (HRA) and UK Medical Research Council (MRC) the research does not need NHS Research Ethics Committee (REC) review for sites in England. The study was declared as 'research'. The HRA decision tool outcome is provided as Supplementary Material Not for Review. The relevant section is:

Will your study involve potential research participants identified in the context of, or in connection with, their past or present use of services (NHS and adult social care), including participants recruited through these services as healthy controls?
This excludes:
* Research where participants have been identified independently of the NHS but because they have a condition that was diagnosed by the NHS (e.g. patients with cancer which may have been diagnosed by the NHS but who are identified from a cancer charity's contact list to be participants in a research project that is otherwise being conducted independently of the NHS).

Data availability statement

The data that support the findings of this study are available from the corresponding author on reasonable request.

Authors contribution

SM wrote the paper. DC was responsible for the data extraction. MVG and JE were responsible for the methods and results sections of AI details and statistical analysis. VP was the experienced grader and provided the retrospective grading analysis. JV and AH were responsible for diabetic retinopathy screening background, failsafe, and logistical considerations for the introduction and discussion. RM provided feedback on the presentation of the results. PK was responsible for checking accuracy and feedback on the paper.