THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Deep learning detection of diabetic retinopathy in Scotland's diabetic eye screening programme

Citation for published version:
Fleming, A, Mellor, J, McGurnaghan, S, Blackbourn, L, Goatman, K, Styles, C, Storkey, AJ, McKeigue, PM & Colhoun, HM 2023, 'Deep learning detection of diabetic retinopathy in Scotland's diabetic eye screening programme', *British Journal of Ophthalmology*. https://doi.org/10.1136/bjo-2023-323395

Digital Object Identifier (DOI):
10.1136/bjo-2023-323395

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
British Journal of Ophthalmology

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 11. Oct. 2023

Deep learning detection of diabetic retinopathy in Scotland's diabetic eye screening programme

Alan D Fleming[1], Joseph Mellor[2], Stuart J McGurnaghan[1,2], Luke A K Blackbourn[1], Keith A Goatman[4], Caroline Styles[5], Amos J Storkey[3], Paul M McKeigue[2], Helen M Colhoun[1], and on behalf of the Scottish Diabetes Research Network Epidemiology Group

[1] The Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK
[2] The Usher Institute, University of Edinburgh, Edinburgh, UK
[3] School of Informatics, University of Edinburgh, Edinburgh, UK
[4] University of Aberdeen, King's College, Aberdeen AB24 3FX
[5] Queen Margaret Hospital, Dunfermline, Fife, UK

Corresponding author

Joe Mellor, Joe.Mellor@ed.ac.uk

Word count

2950

Synopsis

A deep learning-based grader for diabetic retinopathy, trained against screening programme grades, has greater efficacy than iGradingM which has been safely saving manual grading effort in the Scottish Diabetic Eye Service since 2011.

Abstract

Background / Aims: Support vector machine based automated grading (known as iGradingM) has been shown to be safe, cost-effective and robust in the diabetic retinopathy (DR) screening programme in Scotland (DES). It triages screening episodes as gradable with no diabetic retinopathy versus manual grading required. The study aim was to develop a deep learning based autograder (DLAG) using images and gradings from DES and to compare its performance with that of iGradingM.

Methods: Retinal images, quality assurance (QA) data and routine DR grades were obtained from national datasets in 179944 patients for years 2006-2016. QA grades were available for 744 images. We developed a DL-based algorithm to detect whether either eye contained ungradable images or any DR. The sensitivity and specificity were evaluated against consensus QA grades and routine grades.

Results: Images used in QA which were ungradable or with DR were detected by DL with better specificity compared to manual graders and compared to iGradingM at the same sensitivities ($p<0.001$ and $p<0.001$ respectively). Any DR according to DES final grade, was detected with 89.19% (270392/303154) sensitivity and 77.41% (500945/647158) specificity. Observable disease and referable disease were detected with sensitivities of 96.58% (16613/17201) and 98.48% (22600/22948) respectively. Overall 43.84% screening episodes would require manual grading.

Conclusion: A DL-based system for DR grading was evaluated in QA data and in images from 11 years in 50% of people attending a national DR screening programme. The system could reduce the manual grading workload and could perform better than the current automated grading system.

What is already known on this topic; Many studies have shown that deep learning can achieve close to manual grading performance of retinal images obtained in diabetic retinopathy screening.

What this study adds; We have determined that a deep learning-based system can detect screening episodes with disease or which are ungradable at better than manual grading in images selected for quality assurance processes in the screening programme. The system performs well in data covering more than a decade from 50% of people who attended the diabetic retinopathy screening programme in Scotland.

How this study might affect research, practice or policy; If this system were used in NHS Scotland Diabetic Eye Service it might create economic savings by final grading more patient screening episodes than are final graded by the current automated system. The system described in this study may be combined with automated assignment of screening intervals to further increase the efficiency of diabetic eye screening.

1. INTRODUCTION

The aim of a screening programme for diabetic retinopathy (DR) is to find patients with diabetes whose level of DR is such that they would benefit from referral to ophthalmology services for timely treatment of the disease. In Scotland all people with diabetes aged 12 and upwards are offered regular screening in which there is a single-field image taken from each eye in the Diabetes Eye Screening programme (DES) [1]. Since 2011 images are triaged for manual grading using a support vector machine based autograder (iGradingM) trained to classify image sets as ungradable or contains any DR (see Figure 1 for details of the grading process in DES) [2]. iGradingM (Medalytix Ltd) has been shown to safely identify about 50% of the screening episodes not requiring manual grading [3], [4], [5]. This is similar to a 48% reduction in manual grading burden achieved by RetMarker software (Retmarker SA, Portugal) introduced around the same time [6]. Although DL-based systems have been reported to have higher sensitivity and specificity for referable DR, head-to-head comparison of five systems found wide variation in sensitivity (from 60% to 86%) and specificity (from 60% to 84%) for any retinopathy [7]. A narrative review by the UK National Screening Committee found that EyeArt was the only DL-based grader with high-quality evidence on performance in any UK screening programme [8].

Although iGradingM has made worthwhile savings in manual grading workload in Scotland's DES, there is an opportunity for a DL-based system to make further savings in manual grading workload by achieving better specificity while maintaining similar sensitivity for detection of disease.

The aim of this study was to compare the rate of detection of any disease or ungradable images by a DL-based grader against manual grading and the current autograder, iGradingM, using images which had been selected for DES quality assurance (QA) procedures. We also aimed to evaluate the rate of detection of screening episodes with any DR or ungradable images by the DL-based grader in a large dataset from DES covering 11 years. We evaluated the rates of detection across years of the retrospective data.

2. Materials and methods

2.1. Details of the DES operation and grading system

In Scotland the DES operated as follows. All patients registered with diabetes with age 12 years and above in Scotland were invited to attend DES. A single 45 degree macula-centred photograph was taken of each eye, with additional photographs as required for better image quality or view of pathology. In a process termed staged mydriasis, pupil dilation with eye drops was performed only if small pupils caused image quality to be inadequate.

Until 2011, final disease grades, as shown in supplementary table 1, were determined through a three-level manual process, figure 1. All screening episodes that received a grade that indicated presence of disease or ungradable images by level 1 grading were passed to level 2 grading. Episodes receiving a referable grade by level 2 grading were reviewed at level 3 mainly by ophthalmologists. All manual graders contributing to the final grade had passed compulsory nationally administered proficiency testing and were assessed in QA processes. After 2011,

automated grader iGradingM was in operation and reduced the manual grading workload by final grading screening episodes that it deemed as being gradable and with no disease. iGradingM is subject to the same QA procedures as manual grading.

2.2. QA system in DES

DES runs internal and external QA processes ensuring high robustness of the manual grades [9]. In the external QA process, 100 images were selected for each biannual round by an ophthalmologist such that the set was enriched for referable and difficult examples. Each image was assessed by a panel of seven to nine ophthalmologists, blinded to the system grade, to achieve a consensus grading. The QA process evaluates iGradingM and graders qualified to perform level 1 or 2 grading. DES lead clinicians selected the images to be used in QA from the screening programme images.

2.3. Data sets and reference standard

Two data sets were available for this study. The first data set contained images used in external QA of DES. This included the reference grading, iGradingM test results and gradings by graders qualified to grade at levels 1 and 2 and who participated in this QA process between 2011 and 2016. The reference grade was the retinopathy and maculopathy grade and gradability assigned by the largest number of ophthalmologists, the most severe being chosen in the case of ties. The second data set contained routine screening data from DES, including retinal fundus photographs and their final eye-level grades. It included all screening episodes with at least one eye-level final grade that took place between 1st January 2006 and 31st December 2016. Test results from iGradingM were not available in the routine data. The time range was chosen pragmatically; for this time range a single extraction process from the screening programme database (Soarian; Siemens, Germany) was possible.

2.4. Training and tuning of the DLAG

We developed a system to operate as a DLAG as follows. The routine screening data was split into development and testing sets at the person-level. All individuals whose images had ever been used in QA processes plus a random selection of individuals were assigned to the development set. This was then split to training and tuning sets. The proportions of individuals in training, tuning and test sets were 40%, 10% and 50% respectively.

DLAG included two deep neural networks with Resnet-101 architecture [10]. The input to each neural network was the red, green and blue planes of a fundus photograph scaled to 672 by 672 pixels. Both networks were trained using a cross-entropy loss function over 400 epochs. Tuning set accuracy was assessed every 50 epochs. The network weights at the epoch with best tuning accuracy were used in the final network in further analysis. Initial learning rate was 0.1 and a cosine learning rate schedule was used. Fundus images were preprocessed to remove black borders and during training were augmented randomly by rotation, resize, cropping, colour jitter, and gridded cutout. One neural network had a single output and was trained to determine image laterality based on the labelling of the DES images. This was used to determine if images are of the same eye.

The other neural network was trained to determine grade of retinopathy, maculopathy and gradability. A vector of 6 outputs represented retinopathy and image gradability and a vector of 4 outputs represented maculopathy and image gradability. A softmax operation was applied to each vector of outputs. The training loss for this network was sum of the cross-entropy loss for each output vector. Given two vectors of softmax values from two images of the same eye at the same screening episode, [$R0_1$, $R1_1$, $R2_1$, $R3_1$, $R4_1$, U_1] and [$R0_2$, $R1_2$, $R2_2$, $R3_2$, $R4_2$, U_2], a single vector [min($R0_1$,$R0_2$), max($R1_1$,$R1_2$), max($R2_1$,$R2_2$), max($R3_1$,$R3_2$), max($R4_1$,$R4_2$), max(U_1,U_2)] was created, where R0, …, R4 are softmax representing retinopathy grades and U is the softmax representing ungradable. The process was similar for maculopathy.. The resulting vectors from right and left eye images, were concatenated and input to a linear regression model which was trained for classifying a screening episode as any DR or ungradable versus gradable with no disease. Thus, the neural network output vectors from all images in a screening episode, were collapsed to a single value indicating the probability whether the screening episode was any DR or ungradable versus gradable with no disease.

A threshold was chosen at the 80th percentile of this probability in gradable screening episodes with no disease. The tuning set of images was used to train the logistic regression and determine the threshold. Screening episodes with probability greater than or equal to the threshold were classed as test positive and were otherwise classed as test negative. The threshold was chosen to give DLAG a specificity of approximately 80% since this was likely to be much higher than the specificity of iGradingM. In practice a different threshold may be chosen.

Information on whether or not a patient had a missing eye were not available and so screening episodes without a gradable image for either eye were classified as test positive. In practice, this would allow a person to assess the cause of missing images in one eye.

In the test portion of the routine screening data set, the measure of performance was the area under a receiver operator characteristics (ROC) for detecting any disease or ungradable images according to DES final grade. This allowed a comparison of performance with iGradingM. When operating at the threshold described above, the rates of test positive by DLAG were determined for multiple DES final grade disease levels. Workload for manual grading was estimated as the total number of screening episodes which were test positive and which would therefore require attention by manual graders.

2.5. Comparison of performance of DLAG with the iGradingM using QA data

In the QA dataset, a measure of the performance of DLAG was the area under a ROC (AUROC) curve for images where the reference said the image was ungradable or had any disease. For DLAG operating at a given threshold, sensitivity and specificity were used as performance measures. Sensitivity and specificity were also used as performance measures of three comparator grading methods (manual grading by graders participating individually in QA, the final DES grade and iGradingM test result). To compare DLAG to each comparator grading, a threshold was applied to DLAG output so that it had either the same sensitivity or same specificity as each comparator grading. Measures or performance were against the reference grade defined by the panel of ophthalmologists as described above.

2.6. Statistical methods

Confidence intervals (CI) on proportions were calculated using the Wilson score interval for binomial proportions. Confidence bands on ROC curves were determined using bootstrap sampling with replacement maintaining the same number of cases and controls as in the original sample. McNemar's test with continuity correction was used to compare the sensitivities at the same specificity or sensitivities at the same specificity.

3. Results

The test set of routine screening data contained 950213 episodes from 179944 patients and the median (interquartile range) number of screening episodes per patient was 5 (2-8). 5949 of 944363 episodes (0.63%) had images for only 1 eye available. 90.2% (1708674/1894629) of eyes had one image, 9.2% had two and 0.7% had three or more images. Supplementary table 2 shows, for each year, the numbers of screening episodes, the number joining the cohort that year, mean age and the distributions of gender and type of diabetes. Supplementary figure 1 shows the number of screening episodes per year, the rates at which each disease grade occurred each year and the rate of ungradable episodes per year. The number of people screened has increased yearly and there are notable trends in the rates of disease grades. The number of people with no disease has shown a general increase with corresponding decrease in the disease grades. Between 2006 and 2009 the number of people with referable DR showed a sharp decrease of about 24% per year. Between 2009 and 2016 the decrease was more gradual at about 3% per year. The QA data set contained 744 images with consensus grades, of which: 60 were ungradable, 393 were graded R1, 17 were graded R2, 50 were graded R3, 64 were graded R4, 52 were graded M1, and 150 were graded M2.DLAG was thresholded to create classifiers with either the same sensitivity or the same specificity as each of the comparator grading methods, namely, manual graders participating in QA, the final DES grade and iGradingM, for any disease or ungradable images. The sensitivity and specificity of the comparator grading methods are shown in table **1** along with the sensitivity and specificity of DLAG at each threshold. The specificity of DLAG is higher than manual graders ($p<0.001$) and higher than iGradingM ($p<0.001$), at the same sensitivities. A similar statement can be made regarding a comparison of sensitivities at the same specificity. No difference was observed between sensitivities or specificities of DLAG and DES final grade at the same sensitivity or specificity, respectively. This data is also presented in the ROC curve for detection by DLAG of images with any DR or which were ungradable in supplementary figure 2 (AUROC 0.969 (0.958-0.969)).

In the routine screening data set, the AUROC for detection by DLAG of any DR or ungradable screening episodes was 0.912 (0.911-0.912). The image laterality network achieved an accuracy of 99.76% on the routine grading test set with an AUC of 0.999. Table **2** shows the rates at which the software would give a test negative or test positive result for collapsed levels of DES final grade. The proportion of all screening episodes receiving a test positive result was 43.84% (416605/950312) which is the predicted overall manual grading workload. Screening episodes with observable disease, and referable disease were detected with rates 96.58% (16613/17201) and 98.48% (22600/22948). Per-year test positive rates and the per-year test negative rates are shown in Supplementary figure 3. Test positive rates had a decreasing trend

from 2012 to 2016 for most of the disease levels and for ungradable images. There was a corresponding increase in test negative rate for gradable images with no disease over these years. This would result in a decrease in predicted manual grading workload.

***Table 1**: Sensitivity of DLAG at the same specificity as each of three comparator methods and specificity of DLAG at the same sensitivity as each of the three comparator methods, for detection of ungradable image or any disease, in the QA data set. The comparator methods are manual grading by graders participating individually in QA, the final DES grade and iGradingM. Sensitivity and specificity for any disease or ungradable images by each comparator method are also given. Sensitivities and specificities in each row are compared using a p-value from a McNemar test. Note that multiple graders participated in QA and each graded all images. Sens. = sensitivity, spec. = specificity.*

Comparator grading method	Comparator sens. % (n/N)	Comparator spec. % (n/N)	DLAG sens. at same spec. % (n/N)	P (sens.)	DLAG spec. at same sens. % (n/N)	P (spec.)
DES final grade	92.80 (541/583)	90.00 (144/160)	92.97 (542/583)	1.000	90.00 (144/160)	1.000
Gradings by participating graders	95.83 (35443/36985)	75.28 (7120/9458)	96.23 (35589/36985)	0.021	78.12 (7389/9458)	<0.001
iGradingM	92.97 (542/583)	61.88 (99/160)	97.60 (569/583)	<0.001	89.38 (143/160)	<0.001

***Table 2**: Performance of DLAG in screening episodes according to final grade of the screening programme in the test portion of the routine grading set. The test negative and positive rates are given for each collapsed retinopathy and maculopathy grade. The total test positive rate is an estimate of the manual grading workload.*

Final screening programme grade	Grade codes	Number of episodes n (%)	Test negative rate % (n/N)	Test positive rate % (n/N)
No retinopathy	R0, M0	647158 (68.1)	77.41 (500945/647158)	22.59 (146213/647158)
Mild retinopathy	R1, M0	226625 (23.8)	12.73 (28850/226625)	87.27 (197775/226625)
Observable maculopathy	R0-R1, M1	13921 (1.5)	4.15 (578/13921)	95.85 (13343/13921)
Observable background retinopathy	R2, M0-M1	3280 (0.4)	0.30 (10/3280)	99.70 (3270/3280)
Referable maculopathy	R0-R2, M2	15921 (1.7)	1.85 (294/15921)	98.15 (15627/15921)
Referable background retinopathy	R3, M0-M2	4008 (0.4)	0.37 (15/4008)	99.63 (3993/4008)

Final screening programme grade	Grade codes	Number of episodes n (%)	Test negative rate % (n/N)	Test positive rate % (n/N)
Proliferative referable retinopathy	R4, M0-M2	3019 (0.3)	1.29 (39/3019)	98.71 (2980/3019)
Ungradable	R6	36380 (3.8)	8.18 (2976/36380)	91.82 (33404/36380)
Total		950312 (100)	56.16 (533707/950312)	43.84 (416605/950312)

4. DISCUSSION

Automated grading has been running in NHS Scotland DES since 2011 using iGradingM, a class IIa medical device registered with the Medicines and Healthcare products Regulatory Agency (MHRA). iGradingM was developed prior to DL with convolutional neural networks being widely available. Apart from QA procedures, the grading of episodes which were final graded by iGradingM was fully autonomous. An upgrade to the automated grading algorithm could improve the sensitivity, specificity or both for detection of screening episodes requiring human observation.

Using screening episodes in 50% of individuals attending the Scottish DES screening programme over 11 years, this study has demonstrated that detection of screening episodes with any DR or ungradable images can be achieved with 89.19% sensitivity and 77.41% specificity. The proportion of episodes for which manual grading would be required if using DLAG is 43.84%, lower than the 50% reported in [4].

A high level of performance in QA processes relative to manual grading and relative to iGradingM are requirements that would need to be satisfied before any grading system can be deployed in DES. This study has shown that DLAG would achieve specificity 90.0% (95% CI 81.2-98.0) at the same sensitivity as the DES final grade when running in this QA data set, which was enriched with challenging cases. This is well above the specificity of iGradingM 61.88 (99/160) on this data.

Many systems have been developed using DL to detect DR and evaluated in many local or regional diabetic eye screening programmes around the world. For example, systems have been trained, tested or both using images from screening programmes in Thailand [11], India [12], Singapore [13], Portugal [6], China [14], [15], several states of the USA [16], [7], Spain (Andalusia) [17] and England [18]. In some of these studies the automated systems were compared against routine DR gradings provided by certified graders either in the full validation cohort [16], [18] or in part of it [7], [19], [6]. Other studies used a reference grade developed for the study over the entire validation cohort [14], [12], [11], [17]. None of these systems have been applied to images from the Diabetic Eye Screening (DES) in Scotland. All of these studies were evaluated cross-sectionally on a single screening episode from each patient.

Strengths of this study are its evaluation against datasets approved for use in a systematic QA process in a DR screening programme, and the size of one of the data sets used, with nearly 1 million screening episodes, being 50% of an entire national screening programme. In addition the evaluation has been performed across data from 11 years. The final grade from the

screening programme was used as the reference standard for the latter evaluation, and this was obtained mainly from manual graders who were nationally accredited and who took part in regular centrally administered QA processes. Also, since 2011, an automated grading system, iGradingM operated in many centres and made the final grade for screening episodes where it found gradable right and left eye images with no microaneurysms.

In the test using the routine grading set, the system was set to operate at a constant predetermined threshold applied to the predicted probability of having any DR or being ungradable. The threshold was selected on data separate from the test set. Over a sequence of years, however, there were noticeable trends in the test negative and test positive rates of the system. There appears to be an inflexion in these curves around 2011 which could be due to changes in policy at the time. The apparent trade-off between test negative and test positive rates since 2011, suggests more consistent performance might be obtained by regular recalibration of the threshold mentioned above. The optimal method for this recalibration has not been determined and is beyond the scope of this paper.

This study has the following limitations. Firstly, a reference grading was available only at the image level. However these images were chosen for a QA procedure that was designed to ensure the quality of grading in the screening programme and so we considered that performance level of DLAG on these images is an important factor. Secondly, a two-year screening interval policy for patients with two consecutive screening episodes with no retinopathy and no maculopathy was introduced by DES in 2021. The study data pre-dates this important change and we did not attempt to evaluate how automated grading might be affected by it. Thirdly, we did not account for repeated evaluation over multiple screening episodes on each patient. This is representative of a real world setting where patients have regular screening appointments and the current images are graded independently of previous grades. Fourthly, we have not been able to stratify the results according to whether or not mydriasis has been performed. Fifthly, in the routine screening dataset the DES final grade was used as reference though this was not a consensus grading. Some of those grades were based on a single manual grading or iGradingM grading. These had passed systematic QA procedures. Lastly, the system has been evaluated so far only in the context of Scotland's population and screening policy. Notably this is a predominantly white population and usually a single retinal image is taken of each eye.

In conclusion, DLAG has been evaluated in QA data and in 50% of all patients attending a national screening programme over 11 years. It performed at a level similar to manual graders and detected 98.48% of referable disease and 91.82% of ungradable screening episodes, at a specificity of 77.41% for detection of any DR or ungradable images. At this threshold, only 43.84% of screening episodes would have been referred to manual grading. This algorithm represents a potential replacement for the current automated DR grading system currently running in NHS Scotland.

5. Acknowledgements

We thank the Scottish Diabetes Research Network for their role in data generation.

6. Contributorship

AF and JM contributed equally to this paper. AF and JM conceived and designed the study. HC, PM and AS made important contributions to study design. LB, KG and SM were involved in data cleaning, harmonization, quality-control and databasing of the data. JM, AF and KG coded and performed the data analysis methods. AF drafted the initial manuscript. All authors made critically important contributions to manuscript revision. All authors approved the final manuscript. Mike Black and Neville Lee of the Scottish Diabetic Eye Screening Collaborative oversaw database extraction and transfer. HC is guarantor of the overall content.

7. Funding

This work was funded by Juvenile Diabetes Research Foundation (JDRF) grant number 2-SRA-2019-857-S-B.

8. Competing interests

Helen Colhoun is Principal Investigator on the above JDRF grant. The employment of Alan Fleming and Joe Mellor was with this funding. Helen Colhoun and Paul McKeigue have declared stock options in Bayer AG and Roche Pharmaceuticals. Helen Colhoun has received grants from Astra Zeneca LP, Regeneron, Pfizer Inc, Novo Nordisk, Eli Lilly and Company and is on advisory panels or boards of Novo Nordisk, Eli Lilly and Company, Regeneron, Novartis Pharmaceuticals, Bayer AG and Sanofi Aventis. Helen Colhoun has received payments for Speakers Bureaux and Honoraria from Eli Lilly and Company, Regeneron and Novartis Pharmaceuticals. No other authors have declared any competing interests.

9. Ethics statement

This research, and access to the data sets used, was approved by the Public Benefit and Privacy Panel for Health and Social Care (application 1617-0147) and by West of Scotland Research Ethics Committee (reference 21/WS/0047).

10. References

[1] Facey K, Cummins E, Macpherson K, Morris A, Reay L, Slattery J. Health technology assessment report 1: Organisations of services for diabetic retinopathy screening. Glasgow: Health Technology Board for Scotland 2002.

[2] 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. British Journal of Ophthalmology 2007;91:1512–1517.

[3] Black M. The Scottish experience - automated grading. In:, https://www.eyescreening.org.uk/; 2016.

[4] Styles CJ. Introducing automated diabetic retinopathy systems: It's not just about sensitivity and specificity. Eye 2019;33:1357–8.

[5] 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. British Journal of Ophthalmology 2010;94:1606–1610.

[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] Lee AY, Yanagihara RT, Lee CS, Blazes M, Jung HC, Chee YE, et al. Multicenter, head-to-head, real-world validation study of seven automated artificial intelligence diabetic retinopathy screening systems. Diabetes Care 2021;44:1168–75.

[8] Zhelev Z, Peters J, Rogers M, Allen M, Lowe J, Kijauskaite G, et al. Automated grading to replace level 1 graders in the diabetic eye screening programme. UK National Screening Committee 2021.

[9] Goatman K, Philip S, Fleming A, Harvey R, Swa K, Styles C, et al. External quality assurance for image grading in the scottish diabetic retinopathy screening programme. Diabetic Medicine 2012;29:776–83.

[10] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In:. Proceedings of the ieee conference on computer vision and pattern recognition, 2016, pp. 770–8.

[11] Ruamviboonsuk P, Tiwari R, Sayres R, Nganthavee V, Hemarat K, Kongprayoon A, et al. Real-time diabetic retinopathy screening by deep learning in a multisite national screening programme: A prospective interventional cohort study. The Lancet Digital Health 2022;4:e235–44.

[12] 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;316:2402–10.

[13] 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;318:2211–23.

[14] Zhang Y, Shi J, Peng Y, Zhao Z, Zheng Q, Wang Z, et al. Artificial intelligence-enabled screening for diabetic retinopathy: A real-world, multicenter and prospective study. BMJ Open Diabetes Research and Care 2020;8:e001596.

[15] Huang X-M, Yang B-F, Zheng W-L, Liu Q, Xiao F, Ouyang P-W, et al. Cost-effectiveness of artificial intelligence screening for diabetic retinopathy in rural china. BMC Health Services Research 2022;22:1–12.

[16] 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 Technology & Therapeutics 2019;21:635–43.

[17] Jimenez-Carmona S, Alemany-Marquez P, Alvarez-Ramos P, Mayoral E, Aguilar-Diosdado M. Validation of an automated screening system for diabetic retinopathy operating under real clinical conditions. Journal of Clinical Medicine 2021;11:14.

[18] Heydon P, Egan C, Bolter L, Chambers R, Anderson J, Aldington S, et al. Prospective evaluation of an artificial intelligence-enabled algorithm for automated diabetic retinopathy screening of 30 000 patients. British Journal of Ophthalmology 2021;105:723–8.

[19] Li Z, Keel S, Liu C, He Y, Meng W, Scheetz J, 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.

11. FIGURE CAPTIONS

Figure 1. Grading in the DES in Scotland that includes the iGradingM Autograder and up to three levels of manual grading. A possible role of the proposed Deep Learning Autograder would be a direct replacement of iGradingM Autograder.

Supplementary figure 1. Yearly trends from 2006 to 2016 for people included in the test set. Total number is shown in cyan. The prevalence of screening episodes graded as having no disease are shown in blue. The prevalence of screening episodes with each grade of retinopathy and maculopathy are shown in red and green respectively, and the prevalence of ungradable screens is shown in black. bg.=background, mac.=maculopathy.

Supplementary figure 2. Receiver operator characteristics curve with 95% confidence band for detection of images with any DR or ungradable image used in QA. The sensitivity and specificity of the comparator grading methods, the final screening programme grade, manual graders participating in QA and iGradingM are also indicated with bars indicating 95% confidence interval.

Supplementary figure 3. Performance of DLAG in each calendar year. Test negative rate is shown for no DR or ungradable images (equivalent to specificity for any DR or ungradable image). Test positive rate is shown for detection of any DR or ungradable images, and separately for mild background DR, observable disease, referable disease and ungradable images. Bars represent 95% confidence intervals.

12. SUPPLEMENTARY METHODS

The grading protocol in place in DES in Scotland is laid out in supplementary table 3. During the period of the study data, referable maculopathy was defined as exudates or blot haemorrhages within one disc diameter of the centre of the fovea. However, in this study we used the definition of referable maculopathy that was introduced in the DES in 2021 which is the presence of exudates (not blot haemorrhages) within one disc diameter of the centre of the fovea. An eye originally graded as referable maculopathy was changed to no maculopathy unless the eye had been graded as having exudates within one disc diameter of the fovea.

***Table 3**: Screening grades and their coding wit the corresponding patient triaging outcome as used by DES during the study period.*

Retinopathy	Outcome	Maculopathy	Outcome	Gradability	Outcome
No DR anywhere (R0)	12 month rescreen	No maculopathy (M0)	12 month rescreen	Adequately visualised	
Mild background DR (R1)	12 month rescreen	Observable maculopathy DR (M1)	6 month rescreen	Not adequately visualised (R6)	Recall to slit-lamp

Retinopathy	Outcome	Maculopathy	Outcome	Gradability	Outcome
Observable background DR (R2)	6 month rescreen	Referable maculopathy (M2)	Refer to ophthalmology		
Referable background DR (R3)	Refer to ophthalmology				
Proliferative DR (R4)	Refer to ophthalmology				

13. SUPPLEMENTARY RESULTS

***Table 4**: Supplementary table 2. Characteristics of study screening episodes by calendar year. SD=standard deviation; T1DM and T2DM=diabetes mellitus types 1 and 2.*

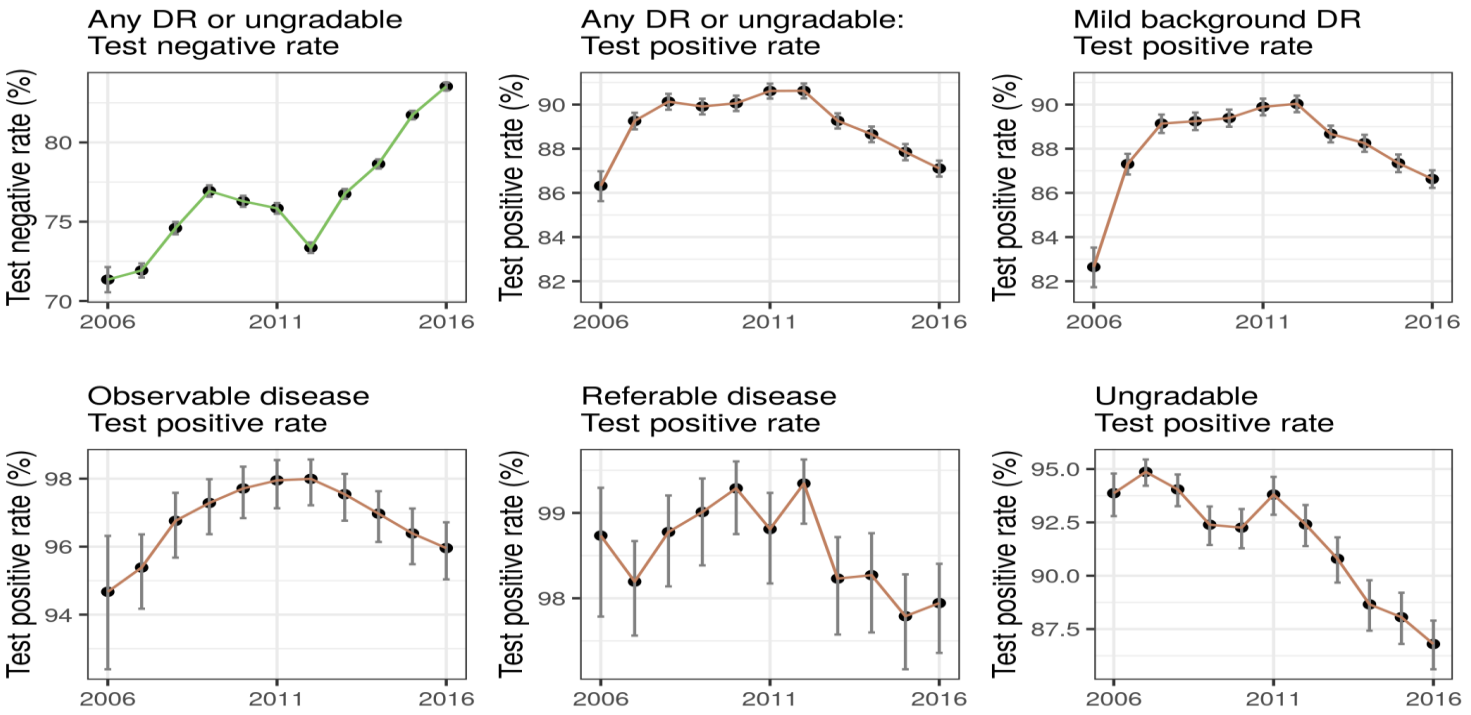
			N	Mean (SD)		% subjects	
Year	Subjects	Screens	First screen	Age	Sex (male)	T1DM	T2DM
2006	22,391	22,616	22,391	60.8 (14.8)	55.71	13.23	84.16
2007	64,581	66,063	49,375	61.2 (14.8)	55.34	12.03	85.49
2008	73,407	74,671	21,741	61.3 (14.6)	55.97	11.04	86.38
2009	78,253	80,365	13,551	61.4 (14.6)	56.40	10.71	86.68
2010	84,479	86,121	11,862	61.6 (14.5)	56.71	10.43	86.84
2011	88,399	90,526	10,240	62 (14.5)	56.75	10.17	87.04
2012	94,363	96,855	10,269	62.2 (14.4)	57.16	10.04	87.16
2013	99,269	102,006	10,830	62.6 (14.3)	57.35	9.70	87.39
2014	103,297	106,793	9,616	62.9 (14.3)	57.34	9.59	87.37
2015	105,526	108,333	9,831	63.2 (14.3)	57.33	9.43	87.41

			N	Mean (SD)			% subjects
Year	Subjects	Screens	First screen	Age	Sex (male)	T1DM	T2DM
2016	111,205	115,864	10,238	63.6 (14.2)	57.58	9.24	87.49
Total	179,944	950,213	179,944	62.3 (14.5)	56.87	10.18	86.95

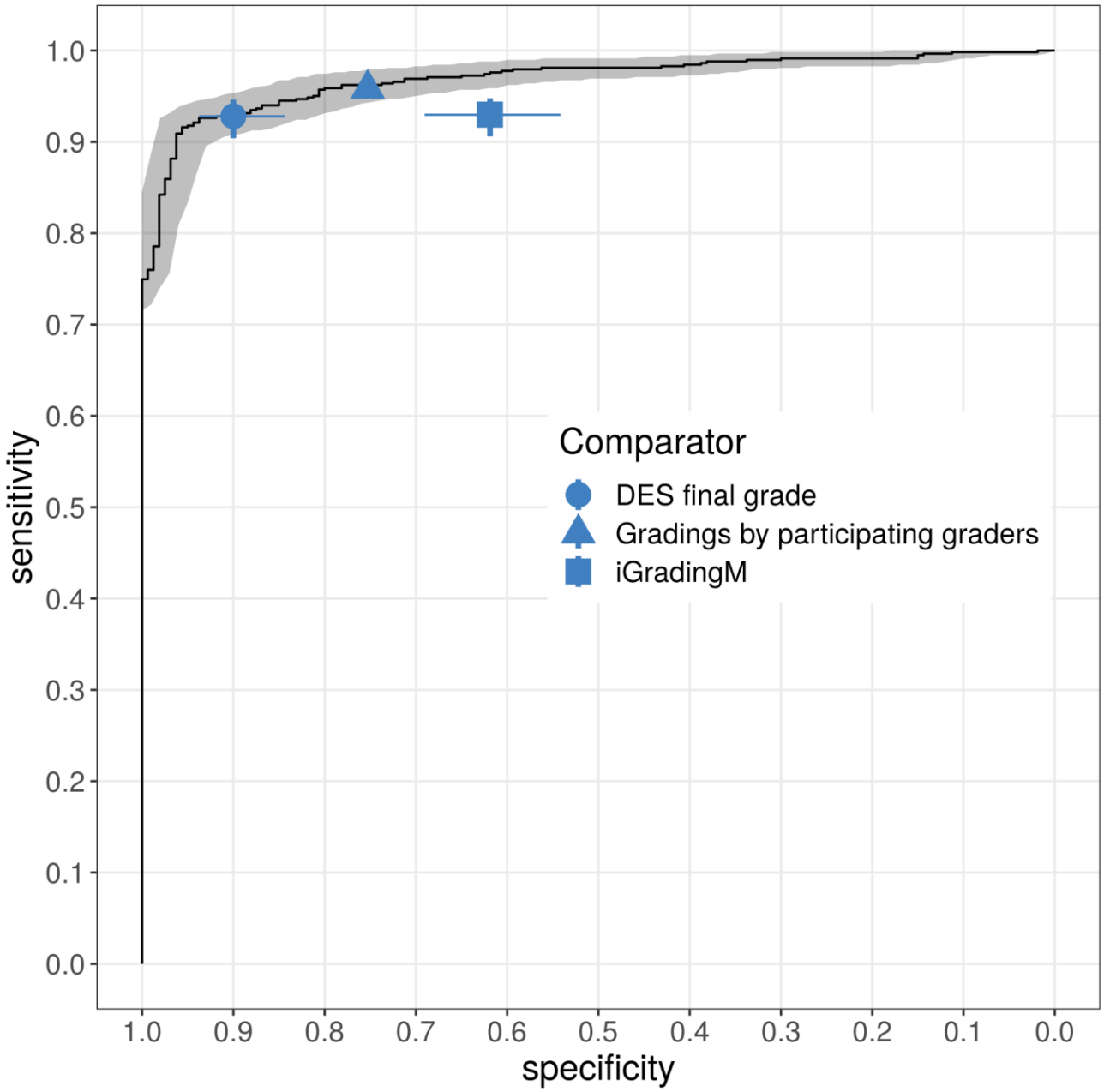
Supplementary figure 1 shows the prevalence of retinopathy and maculopathy grades by calendar year.

Supplementary figure 2 shows receiver operator characteristics curve with confidence band for detection of images used in QA with any DR or which are ungradable. The sensitivity and specificity of the comparator grading methods relative to the majority grade of seven to nine ophthalmologists, namely, manual graders participating in QA, the final screening programme grade and iGradingM, are overlaid. These data can be compared to table 1.

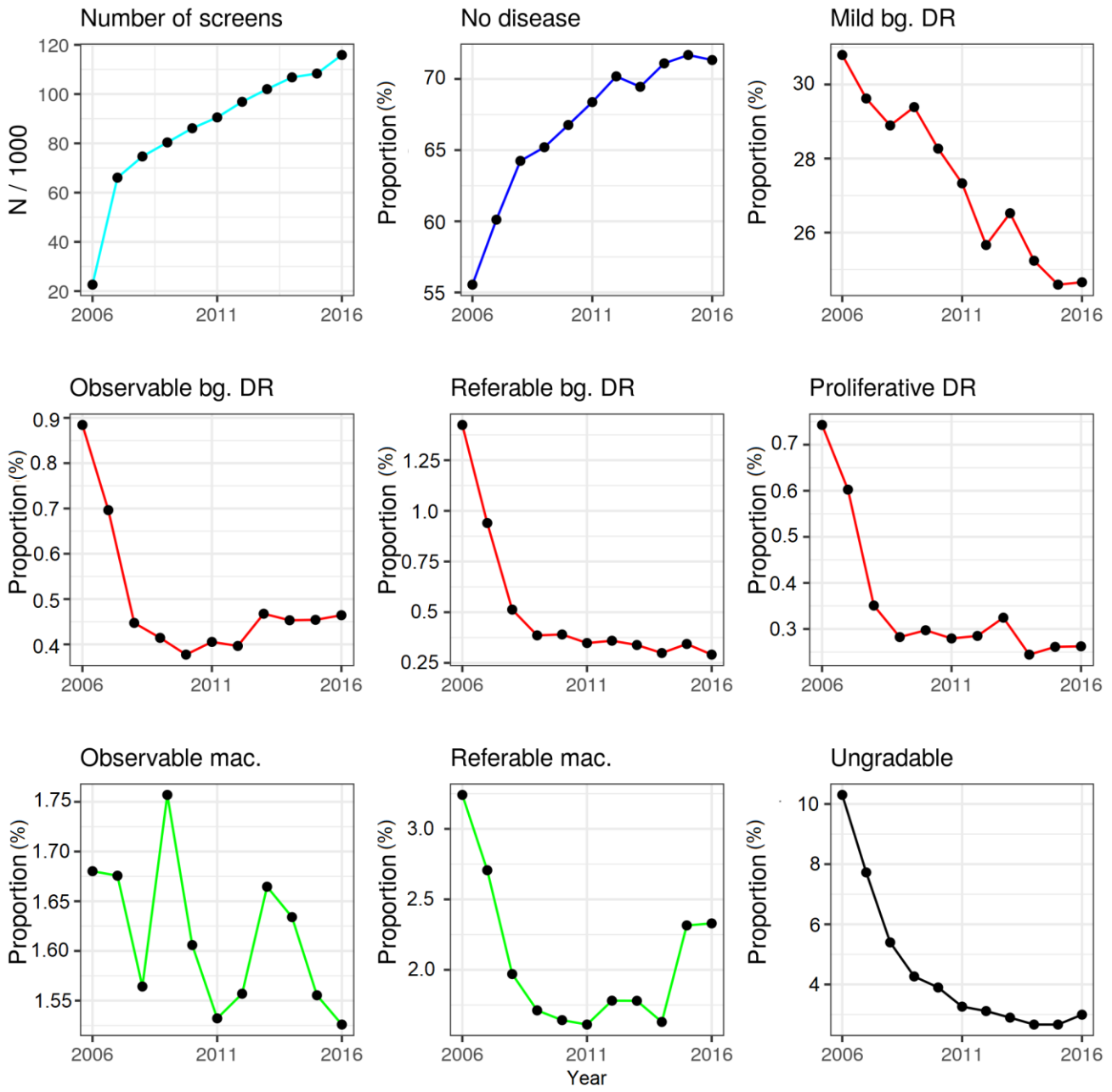
Supplementary figure 3 shows the test negative and test positive rates by calendar year.



Supplementary figure 1. Yearly trends from 2006 to 2016 for people included in the test set. Total number is shown in cyan. The prevalence of screening episodes graded as having no disease are shown in blue. The prevalence of screening episodes with each grade of retinopathy and maculopathy are shown in red and green respectively, and the prevalence of ungradable screens is shown in black. bg.=background, mac.=maculopathy



Supplementary figure 2. Receiver operator characteristics curve with 95% confidence band for detection of images with any DR or ungradable image used in QA. The sensitivity and specificity of the comparator grading methods, the final screening programme grade, manual graders participating in QA and iGradingM are also indicated with bars indicating 95% confidence interval.



Supplementary figure 3. Performance of DLAG in each calendar year. Test negative rate is shown for no DR or ungradable images (equivalent to specificity for any DR or ungradable image). Test positive rate is shown for detection of any DR or ungradable images, and separately for mild background DR, observable disease, referable disease and ungradable images. Bars represent 95% confidence intervals.

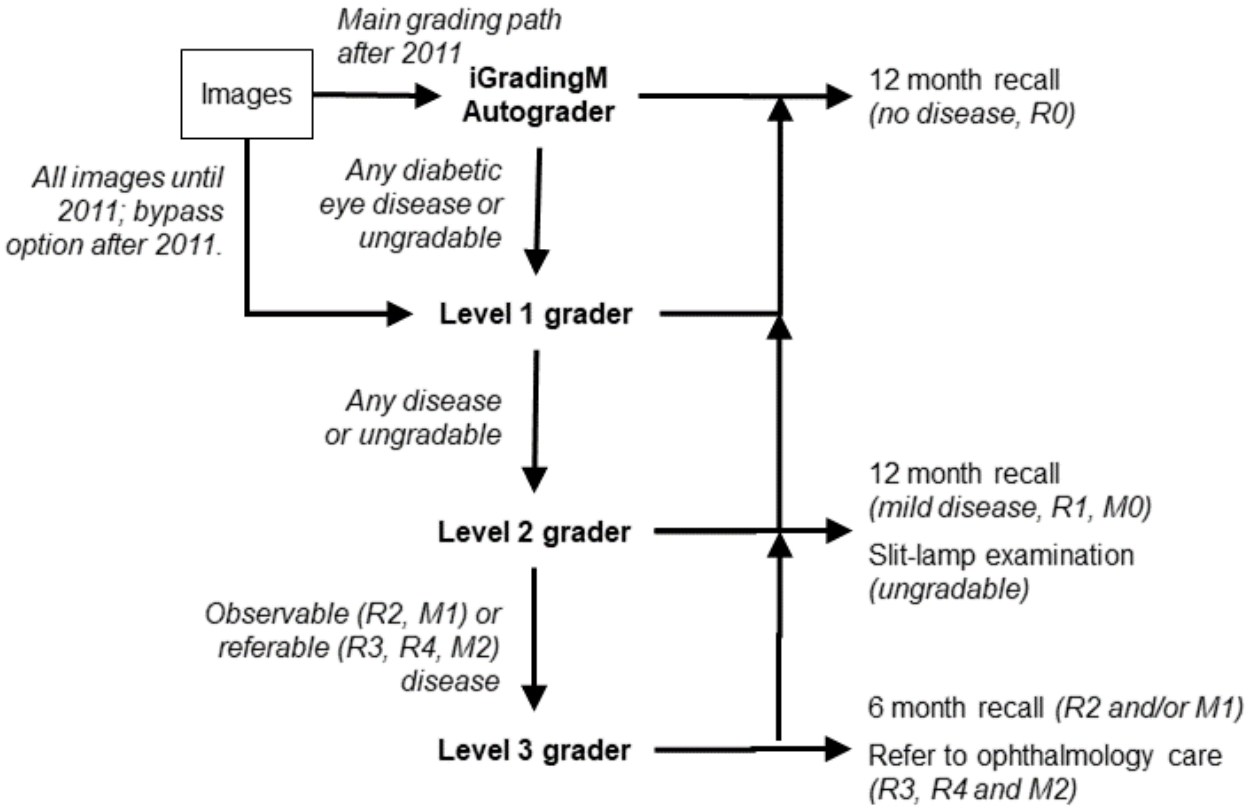


Figure 1. Grading in the DES in Scotland that includes the iGradingM Autograder and up to three levels of manual grading. A possible role of the proposed Deep Learning Autograder would be a direct replacement of iGradingM Autograder.