THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Prediction of retinopathy progression using deep learning on retinal images within the Scottish screening programme

Citation for published version:
Mellor, J, Jiang, WJ, Fleming, A, McGurnaghan, S, Blackbourn, L, Styles, C, Storkey, AJ, McKeigue, PM & Colhoun, HM 2024, 'Prediction of retinopathy progression using deep learning on retinal images within the Scottish screening programme', *British Journal of Ophthalmology*. https://doi.org/10.1136/bjo-2023-323400

Digital Object Identifier (DOI):
10.1136/bjo-2023-323400

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. May. 2024

Prediction of retinopathy progression using deep learning on retinal images within the Scottish screening programme

Joseph Mellor[2], Wenhua Jiang[2], Alan Fleming[1], Stuart J McGurnaghan[1,2], Luke Blackbourn[1], Caroline Styles[5], Amos J Storkey[3], Paul M McKeigue[2], Helen M Colhoun[1,4]

2023-02-09

[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] Department of Public Health, NHS Fife, Kirkcaldy, UK
[5] Queen Margaret Hospital, Dunfermline, Fife, UK

ORCID identifiers

- Paul McKeigue 0000-0002-5217-1034
- Joseph Mellor 0000-0003-1452-887X
- Helen Colhoun 0000-0002-8345-3288
- Alan Fleming 0000-0003-0642-7331
- Stuart J McGurnaghan 0000-0002-3292-4633
- Luke Blackbourn 0000-0003-4234-8040
- Caroline Styles 0000-0002-6515-1032
- Amos J Storkey 0000-0002-8100-506X

Corresponding authors

joe.mellor@ed.ac.uk

Word count: XXXX

Keywords: Diabetes, Retinopathy, Maculopathy, Deep Learning

Abstract

Background/Aims

National guidelines of many countries set screening intervals for diabetic retinopathy (DR) based on grading of the last screening retinal images. We explore the potential of deep learning (DL) on images to predict progression to referable DR beyond DR grading, and the potential impact on assigned screening intervals, within the Scottish screening programme.

Methods

We consider 21346 and 247233 people with T1DM and T2DM respectively each contributing on average 4.8 and 4.4 screening intervals of which 1339 and 4675 intervals concluded with a referable screening episode. Information extracted from fundus images using DL were used to predict referable status at the end of interval and its predictive value in comparison to screening-assigned DR grade was assessed.

Results

The DL predictor increased the AUC in comparison to a predictor using current DR grades from 0.809 to 0.87 for T1DM and from 0.825 to 0.87 for T2DM. Expected sojourn time – the time from becoming referable to being rescreened - was found to be 3.4 (T1DM) and 2.7 (T2DM) weeks less for a DL-derived policy compared to the current recall policy.

Conclusions

We showed that, compared to using the current retinopathy grade, DL of fundus images significantly improves the prediction of incident referable retinopathy before the next screening episode. This can impact screening recall interval policy positively, for example, by reducing the expected time with referable disease for a fixed workload - which we show as an exemplar. Additionally, it could be used to optimise workload for a fixed sojourn time.

Synopsis/Precis

Using deep learning to predict progression to referable retinopathy from fundus images leads to screening policies with shorter expected sojourn time than the current screening policy in Scotland.

What is already known on this topic

Diabetic retinopathy grading schemes, such as those used in the Scottish Diabetic Retinopathy screening programme, grade fundus images on a scale - increasing in severity - ranging from no retinopathy to proliferative retinopathy. Those with more severe diabetic retinopathy grades are more at risk of their diabetic retinopathy increasing to referral or sight threatening over a fixed time period than those with less severe diabetic retinopathy. Deep learning on fundus images has been shown to predict fundus image gradings at human-level, and more recent studies have shown it can predict progression of diabetic retinopathy.

What this study adds

Using a large cohort from the Scottish Diabetic Retinopathy screening programme, this study provides a thorough quantification of the increment in prediction by using Deep Learning on fundus images to predict progression to referable retinopathy, beyond prediction models based on current grading.

How this study might affect research, practice or policy

The study shows how policies based on deep learning on fundus images can identify those most at risk of developing referable changes and reduce their sojourn time by shortening the screening interval, so they are seen within a minimum time of developing sight threatening changes.

1. Introduction

Screening for diabetic retinopathy (DR) using fundus photography is effective in limiting visual impairment caused by DR[1]. Many countries have instituted screening programmes which typically assign some fixed interval between screens that depends on the level of disease present on the photographs[2–4]. Systematic screening for DR has been running in Scotland since 2006.

The choice of screening intervals in most screening programmes has often been a pragmatic compromise between the total workload the system can afford, the desire to minimize the time a person with referable disease is in this state before they are detected and referred (i.e. the sojourn time) and the feasibility for the person with diabetes (a belief that short intervals would be too burdensome but that very long intervals of more than two years might reduce adherence)[5]. In practice the acceptable sojourn time is often not quantified but the incidence of referable disease at next screen in those assigned to a given interval is used as a proxy[6]. Thus the change in the Scottish system to two-yearly screening was made based largely on the low overall incidence (<0.3%) of referable disease at next screen in those with no DR in 2 consecutive screenings.

There is considerable interest in how screening programmes might be altered in a number of ways to gain efficiencies, reduce sojourn times as well as burden on participants. For example one might continue to use fixed intervals based on grades, but change the length of these intervals as was done in Scotland. Or one might have a different interval for a given grade depending on some other stratifying information such as diabetes type. One might move away from fixed intervals completely to a personalized interval aimed at achieving equity in risk of referable disease at the next screen and so on. To this latter end the increment in AUC for prediction of referable disease by including other individual level covariates such HbA1c and blood pressure on top of grades has been evaluated[5,7]. However it is important to evaluate not just the increment in prediction achieved for any given system change but also other aspects such as workload for a given sojourn time distribution or sojourn time distribution for a given workload[8].

In this paper we evaluate the potential of deep learning (DL) on retinal photographs to assign screening intervals both instead of and in addition to the current grading system.

Here we first assess whether DL can improve the prediction of transition to referable retinal grade at next screen if used instead of or additional to the current grading system. We then quantify what the impact would be on the distribution of sojourn time if DL was used to assign the same fixed intervals of 6, 12, and 24 months as used at present and with the total workload of the system held constant at the present level. As an alternative way to summarise impact on

sojourn time we quantify whether those who did have referable disease at next intervals would have been assigned a shorter interval under use of DL than they were actually assigned under the current system.

2. Methods

2.1. Data sources

The study used the Scottish Diabetes Research Network national dataset (SDRN-NDS) [9] that linked all fundus images between October 2005 and March 2017 from the Scottish Diabetic Retinopathy screening (SDRS) programme to a national register of all people with diabetes in Scotland maintained by Scottish Care Information - Diabetes Care (SCI-DC) for primary care data. Data was also linked to Scottish Morbidity Records (SMR) for out- and in-patient records, and to the General Register Office (GRO) for Scotland for death records. Clinical and retinal grading information in the linked register was available from 2003 to 2020.

We also used data from the DDR Lesion Segmentation Dataset [10], which contained fundus images and pixel-level labelling of the images for a number of retinal lesion types including microaneuysms.

2.2. Retinopathy protocol

The photographic protocol used by SDRS specifies a single macula centred fundus photograph from each eye. A variety of non-mydriatic 45 degree fundus cameras were used. Images from each eye were classed as ungradeable or graded as 1 of 5 retinopathy (R) grades (R0 through R4) broadly based on the Early Treatment Diabetic Retinopathy Study (ETDRS) scale [11] and 1 of 3 maculopathy (M) grades (M0, M1, M2) [12]. All manual graders contributing to the final grade had passed compulsory nationally administered proficiency testing and were assessed in QA processes that ensured grading standards were uniform between graders and between grading centres. There is a high false-positive rate of referral for maculopathy (M2) since lesions on 2D photographs lack specificity for macular oedema. For this reason, we defined a referable state in this study as being graded R3, R4, or a composite of M2 and not being subsequently rescreened. Not being rescreened is used as a proxy for confirmation that the condition of the individual was significant enough to require continued referral to an eye clinic and thus removal from the screening programme. During the study period the screening policy assigned people to either immediate referral to an ophthalmology clinic [13] (for grades R3, R4, or M2), 6 month recall (for grades R2 or M1) or 12 month recall (for R0, R1 and M0) based on grading at the most recent screening episode. In 2021 this changed such that those people who had no diabetic eye disease (graded R0 and M0) for 2 consecutive screening episodes were instead screened every 2 years[14,15]. See appendix for detailed definitions of none/mild moderate/referable.

2.3. Study population

The study included all screening intervals in the SDRS programme that started and ended between 1 January 2007 and 1 January 2019. A screening interval starts with a screening episode (interval-start episode) and ends with another screening episode. The latest screening

episode prior to the interval-start episode is referred to as the previous episode. The first screening interval for each individual and screening intervals starting with ungradable fundus images were excluded. Thus each included screening interval has a previous episode and an interval-start episode. The previous episode was used to assign recall intervals of 24 months within the current screening policy. Observed screening intervals were censored at the first referable state, death, or the end of the study period.

2.4. Analysis overview

Analysis was separated into 3 sequential stages: a DL training stage, a generalized linear model (GLM) fitting stage, and a model performance evaluation stage. Predicted scores from DL models trained in the DL training stage were used alongside grading and clinical information to fit models in the GLM fitting stage, and these GLM models were evaluated in the final evaluation stage.

2.5. Deep-learning analysis

Three deep learning models were trained for this analysis with the following functionalities:

1) **ProgressionDL** model: Takes 2 fundus images, one for each eye,at the start of the interval, and predicts both the log hazard rate of referral at the next screening, the sum of the retinopathy grades of the 2 input fundus images and the sum of the maculopathy grades of the 2 input fundus images.

2) **GradingDL** model: Takes a single fundus image and predicts the current retinopathy grade (R0-R4 or ungradable) and maculopathy grade (M0-M2)

3) **LesionDL** model: Takes a single fundus image andpredicts the pixel locations of microaneuysms. The model outputs the number of pixels containing microaneuysms.

The ProgressionDL model used a hybrid ResNet50-ViT network architecture (vit_small_resnet50d_s16_224 from the timm library by [16]). To provide a single prediction from bilateral fundus image inputs, a multiple-instance learning head, as used in [17], was added to the hybrid ResNet50-ViT network immediately after the final global average pooling layer. The multiple-instance learning module used 4 heads each of dimension 128. The multiple-instance learning module was proceeded by 3 linear layers that mapped the output of the MIL layer to each of the 3 outputs; a) a single scalar output corresponding to the log hazard rate of developing referable disease, b) the sum of retinopathy grades for both eyes, c) the sum of maculopathy grades for both eyes. The network structure is illustrated in Figure 1. The network was trained using the final screening programme grades as described in supplementary methods, as are the details of training GradingDL and LesionDL.

2.6. Statistical analysis

GLMs with a complementary log-log link function - to allow for interval censoring - were used to model the transition to referable DR from the interval-start episode. We considered 2 baseline models. The first used only retinopathy and maculopathy grades at the interval-start episode corresponding to the information presently used to determine recall policy. The second used grades from both the interval-start and previous episodes and clinical covariates including

age, diabetes duration, eGFR, HbA1c, BMI, total cholesterol, smoking status, statin use, and hypertensive drug use from the interval-start episode. We applied a log transform to the eGFR covariate within the model to account for skew. For each baseline a model with and without the DL outputs (from ProgressionDL, GradingDL, and LesionDL) was fitted to the tuning dataset for the T1DM and T2DM cohorts separately. A GLM using only the DL outputs, from all 3 DL models, was also fitted using the tuning dataset for each of the T1DM and T2DM cohorts. Predictive performance was evaluated via the AUC, test log-likelihood and the expected information for discrimination using the test dataset. The difference in test log-likelihood between 2 models provides the strength of evidence that one model improved the predictive performance above the other. A difference in test log-likelihood of 6.9 natural log units is asymptotically equivalent to a p value less than 0.005 for comparison of nested models[18]. The difference in expected information for discrimination between 2 models quantifies the size of improvement in predictive performance. It is interpretable without knowing the absolute value of expected information of discrimination of either model, unlike the AUC (e.g. an improvement from an AUC of 0.98 to 0.99 is a larger improvement than from 0.60 to 0.61).

Sojourn time was estimated as the time between incident referable disease and its detection by retinal photography. The information we know is the time between screening examinations and the eye status at these 2 time points, exact time of incident disease was not observed. However if we assume that the rate of referral is constant over each screening interval, we can use the predicted rate of referral from our predictive model to calculate the expected sojourn time, for each interval in which a transition occurs using predictive models of the hazard rate..

We modelled the hazard function as constant over time, equivalent to a Poisson arrival process with an exponential distribution of time to failure. Clinical measures were the nearest recorded value between prior to each patient's interval-start episode within 730 days. Where unavailable, clinical measures were imputed using the *mi* package for R. The current policy assigns screening recalls of 6 months, 1 year, and 2 years determined by the screening programme grades. To demonstrate the consequence of using DL models to determine the current screening interval we calculated the expected sojourn time for the current policy and a policy derived from our DL-only model. The DL-derived policy was constructed such as to share the same proportion of assigned screening recalls as the current policy. Using all screening intervals from the test set the proportion of 6 month, 1 year, and 2 year recalls allocated by the current policy, derived from the grades from the interval-start episode, was determined. The DL-enhanced GLM using grades from the interval-start episode was used to predict the hazard rate for test set screening interval. The intervals were ranked by ascending hazard rate. The DL policy was to allocate the top ranking intervals a recall of 6 months, and the bottom ranking intervals a recall of 2 years with the remaining intervals allocated a recall of 1 year such that the proportion of recalls was the same as for the current policy. The expected sojourn time over a 2 year window was calculated for each policy given the DL-predicted hazard rates of intervals ending in referral. When calculating the expected sojourn time it was assumed that conditional on not being referred within an allocated recall interval, the individual would then be assigned the same recall again until either they were referred or the 2-year window elapsed. Using the same screening intervals considered for the calculation of expected sojourn time, the mean recall interval length of those people who were observed to transition to referral at the next screening episode was calculated for each considered policy. A smaller mean recall interval for those transitioning is indicative of a preferable policy.

3. Results

3.1. Cohort

Cohorts included 21346 and 247233 people with T1DM and T2DM respectively. Both cohorts were divided into 3 sets: a DL training set (8984 T1DM and 119702 T2DM individuals); a tuning and GLM-fitting set (4944 T1DM and 51012 T2DM individuals); and a test set (7418 T1DM and 76519 T2DM individuals). In the T1DM cohort, of the 1517 screenings with a provided grade of M2, but that returned to the screening programme, 1166 (76.9%) returned within 2 years and for T2DM of the 4569 screenings with a provided grade of M2, but that returned to the screening programme, 3730 (81.6%) returned within 2 years. Clinical characteristics are given in Table **1**.

3.2. Do DL-based models predict transition to referable DR more accurately than screening programme grades?

As shown in Table **2** the GLM using only DL outputs increased the AUC, in comparison to a GLM using grades from the interval-start episode, from 0.809 to 0.871 for T1DM (an increase in expected information for discrimination of 0.7 bits and an increase in test log-likelihood of 203.1 natural log units) and 0.825 to 0.886 (an increase in expected information for discrimination of 0.6 bits and an increase in test log-likelihood of 724.8 natural log units) for T2DM. A GLM with both DL outputs and grades did not improve performance substantially above that achieved using DL alone. Nor did a GLM using DL outputs, grades from previous episode and interval-start, and available clinical covariates. The predictive performance of the GLM using only DL outputs remained high across strata of both age and sex. The predictive performance was reduced conditioned on having a maximum retinopathy grade of R2 at interval-start. A summary of stratified predictive performance can be found in the appendix. When selecting a threshold of the DL-only GLM score using Youden's J statistic for predicting referable disease at next screening: 7565 of 32867 intervals in T1DM and 34748 of 289641 in T2DM that end with a non-referable assessment were predicted to be referable, and 342 of 421 intervals in T1DM and 1019 of 1338 in T2DM that end with a referable assessment were predicted to be referable.

3.3. Is expected sojourn time reduced in a recall policy using DL?

There was a decrease in the estimated expected sojourn time using a GLM based on DL outputs compared to a GLM based on screening programme grades by 3.4 weeks in T1DM and 2.7 weeks in T2DM, as shown in Table 3. Figure 2 shows the distributions of expected sojourn times for screening intervals ending in referral for the DL-only and current policy (based on screening programme grades) for both T1DM and T2DM.

3.4. Are more people who become referable assigned to shorter recall intervals using a DL-derived policy?

The workload-matched DL-derived policies, both derived from our GLM model using interval-start grades and DL outputs and the DL-only GLM, led to more people who were observed to

progress to referral being assigned to a 6 months recall than the current policy and less people observed to progress to referral being assigned to a 2 year recall in both T1DM and T2DM cohorts. This is shown in Table 4. The mean recall interval length for those referable at the next screening episode was 11.2 months (T1DM) and 12.6 months (T2DM) for the current policy. This reduced to 9.7 months (T1DM) and 11.4 months (T2DM) for DL-only GLM-derived policy and 9.7 months (T1DM) and 11.3 months (T2DM) for the DL and grades GLM-derived policy.

4. Discussion

4.1. Statement of principle findings

We have shown that DL applied to fundus images can be used to significantly improve prediction of progression to referable retinopathy beyond the information available from retinopathy grading. As an illustration of likely clinical benefit we estimated expected sojourn time of the current policy and a workload-matched DL-derived policy. This showed that prediction by use of DL could lead to a reduction in the delay in detecting people who have progressed to referable disease. Expected sojourn time could only be estimated using analytical techniques. Therefore we also compared the same policies using mean screening period, and by comparing recall interval allocation distributions, of screening intervals which ended in people presenting to screening with referable disease. In both cases an improvement is demonstrated using DL-based screening interval assignment without increasing the number of screening episodes. We found that a model using only DL outputs was not substantially improved by including the manual grading information, nor by further including additional clinical covariates, which suggests that it would be possible to fully automate the recall policy within the screening programme.

4.2. Comparison with other studies

It is well known that fundus images contain information that is predictive of future DR state and this forms the basis of DR screening programme recall policies. However little is known with about the added value that DL could bring to recall interval assignment compared to current grading systems. It has been demonstrated that DL can predict progression of 2 or more ETDRS grades from fundus images [19], however the cohort size was small and images were from 7-field photography in clinical trial participants with macula oedema and hence are not representative of a screening programme population. Other studies have considered prediction of progression of DR using DL but did not quantify the improvement above current grading[20,21]. With respect to screening policy, a number of approaches have been proposed to improve screening programmes. For example, existing programmes have extended screening intervals to 2 years for those people deemed at lowest risk, supported by evidence of low incidence of referable disease at new screen [6]. Others have proposed using personalised screening intervals based on prediction models that include covariates such HbA1c and blood pressure on top of retinopathy grades, where interval lengths are set to maintain rates of referal per screen[7]. In this paper, similarly to a previous study[8], we consider the explicit trade off between the time a person is in a referable disease state before they are detected and referred (i.e. sojourn time) and the total workload of the system

4.3. Strengths and limitations

A significant strength of the study is that it includes the full population of 268579 patients in a nationwide systematic screening programme for DR. The study had some limitations. Firstly, the referable outcome was based on the screening programme grade. Confirmation of this grade could have been more reliable information from the ophthalmology clinic. This information was not available in this study. However M2 referrals who came straight back into the screening programme after referral were not labelled as referable to reduce false positives. Secondly, because date of incident disease is not known, it is difficult to precisely assess a model which assigned a shorter screening interval than the screening programme. For instance, where the screening programme assigned a 12 month recall interval, which ended with detected referable disease, and the DL model assigned a 6 month recall, we do not know if referable disease would have been present at the hypothetical screening episode after 6 months. We therefore assumed that these patients had high enough risk of progression to justify recall at 6 months. Finally, our findings have not been externally validated.

4.4. Summary

We have shown that DL prediction of progression to referable retinopathy using fundus images can be used to improve screening recall interval allocation within the Scottish Diabetic Retinopathy screening programme. Further validation of our DL score is required before it could be used in practice within the Scottish screening programme. This would include validation against adjudicated grades as opposed to the programme final grade and a validation using more recent screening data.

5. Acknowledgements

We thank the Scottish Diabetes Research Network for the role in data generation. This work was supported by JDRF [grant 2-SRA-2019-857-S-B].

6. Author Contributions

JM conceived and designed the study. HC, PM and AS made important contributions to study design. SM and LB were involved in the cleaning up, harmonization, quality-control and databasing of data in Scotland. JM and WJ performed the analyses. JM developed data analysis methods. JM, WJ, and AF contributed to code preparation. JM and WJ drafted the initial manuscript. All authors made critically important contributions to manuscript revision. All authors approved the final manuscript.

7. Competing interest

Helen Colhoun is Principal Invsetigator on the above JDRF grant. The employment of Joe Mellor, Alan Fleming, and Wenhua Jiang 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.

8. Ethics Statement

This study was approved by the SDRN-Epi steering group, using a data source which is available with the approval of the Public Benefit and Privacy Panel for Health and Social Care (https://www.informationgovernance.scot.nhs.uk/pbpphsc/ (application reference 1617–0147) and by the West of Scotland REC 4 Research Ethics Committee (Ref. 21/WS/0047).

9. References

[1] Jones S, Edwards R. Diabetic retinopathy screening: A systematic review of the economic evidence. Diabetic Medicine 2010;27:249–56.

[2] Ólafsdóttir E, Stefánsson E. Biennial eye screening in patients with diabetes without retinopathy: 10-year experience. British Journal of Ophthalmology 2007;91:1599–601.

[3] Scanlon PH. The english national screening programme for diabetic retinopathy 2003–2016. Acta Diabetologica 2017;54:515–25.

[4] Leese GP, Morris A, Swaminathan K, et al. Implementation of national diabetes retinal screening programme is associated with a lower proportion of patients referred to ophthalmology. Diabetic Medicine 2005;22:1112–5.

[5] Smith JJ, Wright DM, Stratton IM, et al. Testing the performance of risk prediction models to determine progression to referable diabetic retinopathy in an irish type 2 diabetes cohort. British Journal of Ophthalmology 2022;106:1051–6. doi:10.1136/bjophthalmol-2020-318570.

[6] Committee UNS. Screening for diabetic retinopathy 19 november 2015 2015.

[7] Heijden AA van der, Nijpels G, Badloe F, et al. Prediction models for development of retinopathy in people with type 2 diabetes: Systematic review and external validation in a dutch primary care setting. Diabetologia 2020;63:1110–9.

[8] Ochs A, McGurnaghan S, Black MW, et al. Use of personalised risk-based screening schedules to optimise workload and sojourn time in screening programmes for diabetic retinopathy: A retrospective cohort study. PLoS Medicine 2019;16:e1002945.

[9] McGurnaghan SJ, Blackbourn LAK, Caparrotta TM, et al. Cohort profile: The scottish diabetes research network national diabetes cohort a population-based cohort of people with diabetes in scotland. BMJ Open 2022;12. doi:10.1136/bmjopen-2022-063046.

[10] Li T, Gao Y, Wang K, et al. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. Information Sciences 2019;501:511–22. doi:https://doi.org/10.1016/j.ins.2019.06.011.

[11] Group ETDRSR. Early treatment diabetic retinopathy study design and baseline patient characteristics. Ophthalmology 1991;98:741–56.

[12] Collaborative SDRS. Scottish diabetic retinopathy grading scheme 2007 v1.1. 2007.

[13] Looker H, Nyangoma S, Cromie D, et al. Rates of referable eye disease in the scottish national diabetic retinopathy screening programme. British Journal of Ophthalmology 2014;98:790–5.

[14] Leese GP, Stratton IM, Land M, et al. Progression of diabetes retinal status within community screening programs and potential implications for screening intervals. Diabetes Care 2015;38:488–94.

[15] Stratton IM, Aldington SJ, Taylor DJ, et al. A simple risk stratification for time to development of sight-threatening diabetic retinopathy. Diabetes Care 2013;36:580–5.

[16] Wightman R. PyTorch image models. GitHub Repository 2019. doi:10.5281/zenodo.4414861.

[17] Ilse M, Tomczak J, Welling M. Attention-based deep multiple instance learning. In: Dy J, Krause A, editors. Proceedings of the 35th international conference on machine learning, vol. 80, PMLR; 2018, pp. 2127–36.

[18] Stone M. An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. Journal of the Royal Statistical Society: Series B (Methodological) 1977;39:44–7. doi:10.1111/j.2517- 6161.1977.tb01603.x.

[19] Arcadu F, Benmansour F, Maunz A, et al. Deep learning algorithm predicts diabetic retinopathy progression in individual patients. NPJ Digital Medicine 2019;2:1–9.

[20] Bora A, Balasubramanian S, Babenko B, et al. Predicting the risk of developing diabetic retinopathy using deep learning. The Lancet Digital Health 2021;3:e10–9.

[21] Nderitu P, Rio JN do, Webster L, et al. Predicting progression to referable diabetic retinopathy from retinal images and screening data using deep learning. Investigative Ophthalmology & Visual Science 2022;63:2087–F0076.

10. Figure Captions

Figure 1. Diagram vizualising the structure of the proposed ProgressionDL model. A single hybrid ResNet50 ViT is used to process both left and right fundus images. The hybrid ResNet50 ViT outputs are then input into a Multiple Instance Learning module. Those outputs are then input to a linear layer to determine the log hazard rate.

Figure 2. Distributions of sojourn time of screening intervals in the test set observed to transition to referable disease for the current policy and the DL-only policy. The left plot shows comparisons of sojourn distributions for T1DM and the right show them for T2DM. The mean sojourn times are shown as vertical dashed lines. In both plots the density for sojourn times lower than 200 days was larger for the DL-only policy. This shows a larger proportion of people requiring referral would be referred sooner under the DL-only policy than the current policy which in practice would lead to their treatment starting earlier,

Table 1: *Cohort characteristics at interval-start for each cohort stratified by DR status (referable or not) at end of interval. Continuous variables show median and quartile range, and categorical variables show percentages.*

	T1DM		T2DM	
Characteristics	Not referable (89577)[1]	Referable (1339)	Not referable (904576)	Referable (4675)
Age at diagnosis (years)	21.1 (2.4, 58.2)	16 (1.7, 54.5)	57.8 (33.7, 78.2)	53 (27.1, 78.7)
Female (%)	47.44	43.47	42.9	43.42
Diabetes duration at screening (years)	15.8 (2.6, 47)	19 (7.4, 47.7)	7.7 (2.2, 22.4)	12.5 (2.5, 28.2)
Body mass index	26 (17.8, 52.8)	26.1 (18.8, 53.1)	31.9 (22.6, 54.8)	31.9 (21.9, 54.2)
Systolic blood pressure	128 (98, 164)	128 (98, 164.8)	134 (106, 170)	136 (106, 180)
Diastolic blood pressure	75 (55, 110.1)	75 (56, 108.2)	78 (60, 100)	79 (58, 102)
Height	1.69 (1.48, 1.88)	1.7 (1.5, 1.88)	1.68 (1.5, 1.86)	1.68 (1.49, 1.87)
Weight	73.4 (41.2, 111)	73.6 (48.9, 110.5)	89 (59, 137.3)	88.4 (57.6, 137)
HbA1c (mmol/mmol)	68 (42, 115)	81 (50, 130)	54 (37, 103.1)	67 (39, 120)
Total cholesterol (mmol/l)	4.6 (3, 7.1)	4.8 (3.1, 7.4)	4.4 (2.8, 7.1)	4.4 (2.7, 7.4)
eGFR (ml/min/1.73m2)	97.8 (53.7, 149.2)	100.4 (47.8, 140.8)	77.7 (39.3, 112.2)	79 (34.1, 115.7)
Ever smoker (%)	62.44	68.48	72.83	72.6
Statins (%)	1.02	2.32	9.22	4.66
Hypertensive drugs (%)	1.04	1.94	9.47	5.07
Normal vision (%)[b]	99.43	99.4	99.07	98.46
Prior CVD event (%)	4.52	7.62	20.92	23.61

[a] Number of referable and not referable are the number of intervals that had that status at the end of the interval, not the number of patients.

[b]Normal vision is defined as 6/18 or better on the Snellen scale.

Table 2: *Prediction of progression to referability for both cohorts. Models are compared using the AUC, expected information for discrimination (Λ), and change in test log-likelihood from the grades-only GLM (ΔLL).*

	T1DM			T2DM		
Model	AUC	Λ(bits)	ΔLL(nat log units)	AUC	Λ(bits)	ΔLL(nat log units)
Grades model	0.809	0.92	0.0	0.825	1.30	0.0
Grades model + clinical covariates	0.820	1.00	23.9	0.842	1.37	97.0
DL	0.870	1.62	203.1	0.886	1.90	724.8
Grades model + DL	0.871	1.60	202.6	0.887	1.96	776.2
Grades model + DL + clinical covariates	0.873	1.69	225.7	0.883	1.98	764.4

Table 3: *Expected sojourn time under the current policy and the proposed DL-informed policy.*

	Expected sojourn time (weeks)[1]		
Cohort	Current policy	Grade+DL policy	DL policy[2]
T1DM	24.6	21.2	21.2
T2DM	27.6	24.7	24.9

[1] The expected sojourn time within a 24-month window is estimated for each policy given the estimated hazard rates from the DL predictor and the interval length assigned by the policy.

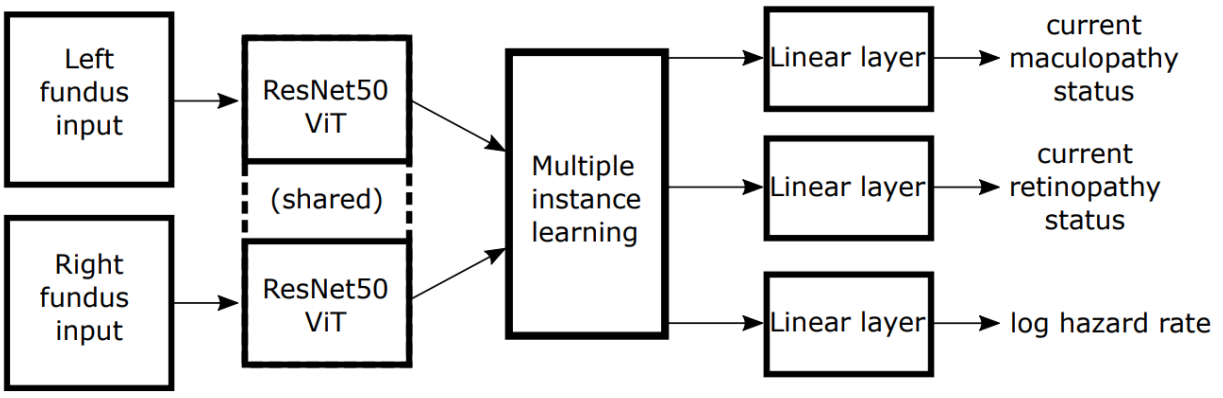
[2]The DL policy ranks screening episodes by the hazard rate predicted by the DL predictor. Screening interval durations for the next episode are assigned in the same proportion to the current policy, with 6-month intervals assigned to those episodes with highest predicted risk and 24-month intervals for episodes with lowest predicted risk.

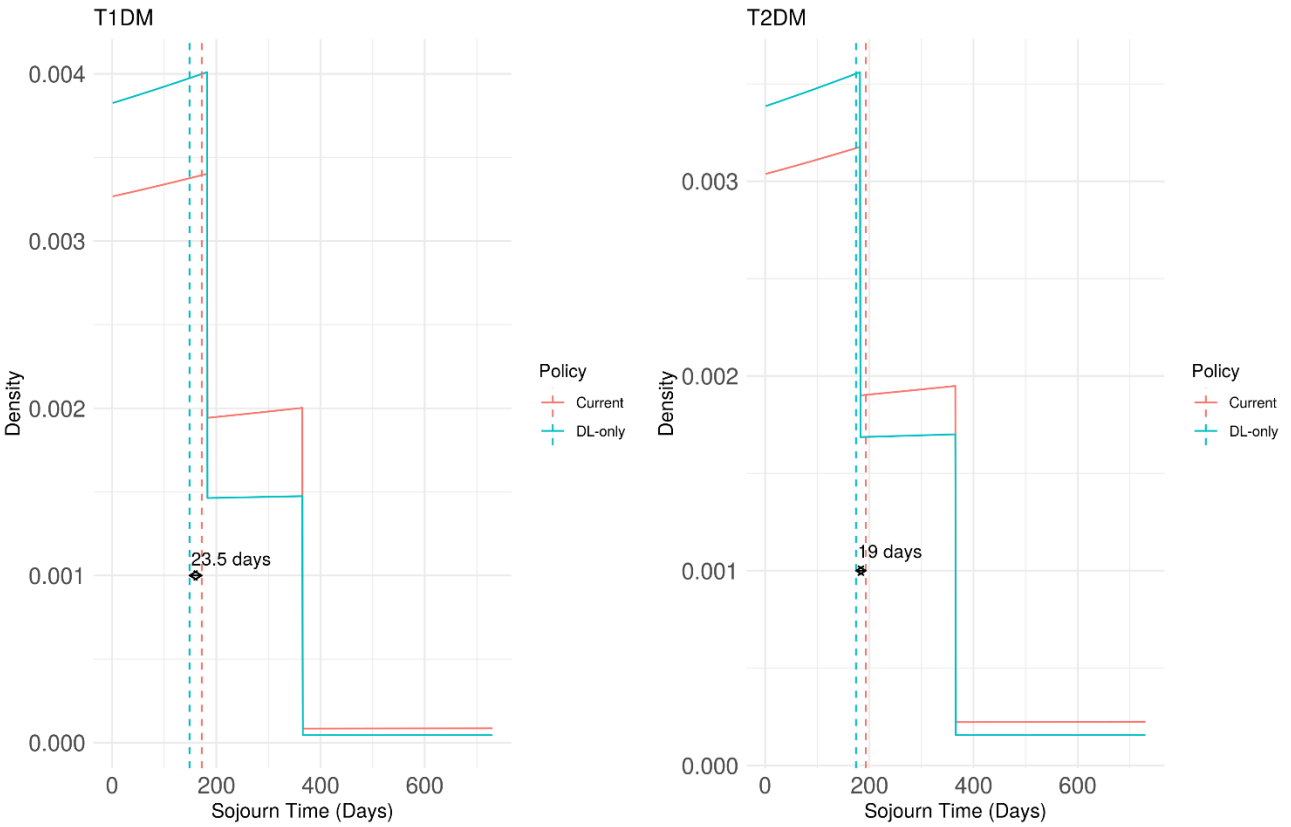
Table 4: *Do DL-based policies assign more screening episodes, where the next screening episode was observed to be referable, to shorter recall intervals? The table shows the percentage of total intervals that ended in referral assigned to each recall interval for each policy.*

	T1DM (421 referrals)			T2DM (1338 referrals)		
Policy interval	Current[1]	Grade+DL[2]	DL[2]	Current	Grade+DL	DL
6 months	25.9% (109)	45.4% (191)	44.9% (189)	22.4% (300)	34.2% (458)	32.7% (438)
1 year	67.9% (286)	51.1% (215)	51.8% (218)	61.4% (821)	54.6% (730)	55.9% (748)
2 years	6.2% (26)	3.6% (15)	3.3% (14)	16.2% (217)	11.2% (150)	11.4% (152)

1 For each observed screening interval in the test set we assign a 6 month, 1 year, or 2 year recall interval as determined by the episode grades in line with current policy.

2 We also assign alternative recall intervals based on a) the DL-only GLM and b) the DL+Grading GLM such that number of 6 month, 1 year, and 2 year recall intervals are the same as for the current policy.




T1DM
0.004
0.003
0.002
0.001
0.000
Density
23.5 days
0
200
400
600
Sojourn Time (Days)
Policy
Current
DL-only
T2DM
0.003
0.002
0.001
0.000
Density
19 days
0
200
400
600
Sojourn Time (Days)
Policy
Current
DL-only