Predicting the future development of diabetic retinopathy using a deep learning algorithm for the analysis of non-invasive retinal imaging

Rom Yovel [1], Aviv Rachelle, [1], Ianchulev Sean [1, 2], Dvey-Aharon Zack [1]

[1] AEYE Health Inc.
[2] New York Eye and Ear, Mount Sinai Hospital, NY

Abstract

Diabetic retinopathy (DR) is the most common cause of vision loss in the working age. While over 90% of sight-threatening cases may be treated if detected early, prevalence of yearly detective screening is low until advanced presentation of the disease. We developed a machine learning algorithm for the prediction of future DR development using fundus photography of otherwise healthy eyes. Our algorithm achieves 0.81 Area Under Receiver Operating Curve (AUC) averaging scores from multiple images on the task of predicting development of referrable DR, and 0.76 AUC when using a single image. In conclusion, risk of DR may be predicted from fundus photography alone. Prediction of personalized risk of DR may become key in treatment and contribute to patient compliance across the board. Further prospective research is necessary.

1. Introduction

Diabetic retinopathy (DR) is a common retinal vascular complication of diabetes mellitus, which is characterized by retinal micro-aneurisms, hemorrhages, neovascularization, and edema in the retina. Diabetic retinopathy can advance to blindness, and is the leading cause of vision-loss at the working age.[1] While over 80% of diabetics develop retinopathy of some degree after 20 years,[2] more than 90% of the sight-threatening cases can be treated, if found early, in time to prevent loss of sight.[3]

Current public health guidelines for individuals with diabetes prescribe annual or bi-annual screening for the presence of DR.[4,5] Multiple clinical studies have demonstrated that screening can lead to early detection and timely treatment which ultimately can prevent serious visual impairment and blindness.[6,7] While retinal screening is essential for patients with diabetes, it requires a specialized eye exam which is often inaccessible for patients. A large percentages of individuals with diabetes forego screening retinal exams and present late in the course of the disease.[8–11] Early intervention is the key to mitigation of DR risk factors and damage. As such, early detection is the most promising way of mitigating the damages of DR.

More recently, AI and machine learning have been applied, with success, to the autonomous diagnosis of referrable (more than mild) DR. There is an FDA approved AI system with sensitivity of 87% and specificity of 90%.[12] More recently, our team reported results of a Pivotal FDA study with 93% sensitivity and 91% specificity for referrable DR on images obtained by a desktop device and 92% and 94% sensitivity and specificity, respectively, on images obtained by a portable camera.[13] Additionally, our team published strong efficacy for DR detection using a portable camera on a separate dataset.[14]

While diabetic retinopathy can be diagnosed by physicians and machine learning systems, there have been no reliable prognostic models to predict DR conversion to date. Some work has been done on finding risk factors for DR, using patient data such as age, HbA1c levels, gender, duration of disease, and so on.[15,16]

Recent work has shown that otherwise “normal” fundus images can be quite informative and predictive when presented to a machine learning algorithm. AI algorithms can interpret subclinical information of the retinal anatomy and make predictions about diseases, even those unrelated to the eye - such as chronic kidney disease (CKD),[17] diabetes,[18] and cardiac arrest.[19] Furthermore, machine learning algorithms have

been trained to predict gender information with high accuracy from mere fundus photography – something previously unattainable with the standard clinical exam.

In this study we present a first-in-class machine learning algorithm which can predict the development of future DR from otherwise normal retinal anatomy.

2. Methods

2.1 Dataset

This research utilized a dataset compiled and provided by EyePACS (http://www.eyepacs.org), comprised of fundus retinal images and expert readings of said images. The data consisted of 156,363 images from 21,730 patients who visited the clinics at least twice between 2016 and 2021. 19.6% of visit pairs were ≤ 12 months apart, 55% were 12-24 months apart, 19.8% were 24-36 months apart, and 5.5% of visits were ≥ 36 months apart (for exact distribution see appendix A). Of the patients, 37% were male and 63% were female or other; mean age was 55 years old (for full demographic data see Table 1).

The dataset contained up to 6 images per patient visit: one macula centered image, one disk centered image, and one center image, per eye. Each eye was graded individually by an expert ophthalmologist for the presence and severity of diabetic retinopathy (DR). DR severity (none, mild, moderate, severe, or proliferative) was graded according to the International Clinical Diabetic Retinopathy scale.[20,21]

The image categorization in the current research was simplified to three severity categories by combining categories 3 to 5 into "more-than-mild DR", as only these levels usually necessitate referral to an ophthalmologist and/or medical and surgical management.[22,23]

	Training set	Validation set
Number of patients	19,531	2,199
Number of images	140,614	15,749
Age: mean, years (s.d.)	55.15 (10.68), n=19495	55.34 (10.40), n=2195
Gender (% male)	0.37, n=19102	0.37, n=2155
HbA1c: mean, % (s.d.)	7.98 (2.26), n=15672	7.99 (2.75), n=1756
Disease duration: mean, years (s.d.)	7.46 (6.44), n=18803	7.21 (6.20), n=2108
Ethnicity		
	62.1% Latin American, 11.4% ethnicity not specified, 9.1% African Descent, 7.4% Caucasian, 5.4% Asian, 2.8% Indian subcontinent origin, 1.1% Other, n=19076	61.4% Latin American, 10.7% ethnicity not specified, 9.7% African Descent, 7.8% Caucasian, 5.6% Asian, 3.1% Indian subcontinent origin, 1.1% Other, n=2147

Table 1: Describes key characteristics of the dataset. The patients were designated randomly to the training set and the validation set.

In order to prepare the dataset for model training and validation, each image was labeled by the maximal DR rating the patient was diagnosed with in a time period following the visit. Towards this purpose each patient visit was rated twice on the DR scale, once for each of the patient's two eyes. Pairs were then created consisting of all possible pairings of each patient's visits in a given time period. Values of the pairings were calculated by measuring the difference in DR ratings, and then taking the maximum value. Each time point (visit) was then assigned the highest value from the pairings in which that timepoint was the first, and each image was labeled by that value. Negative differences were disregarded, as the

regression cause was unknown: true disease regression, clinical intervention, or misdiagnosis. Models were created for each of the chosen time periods.

For instance, a given patient (P) has visited a clinic n times, once a year: v_1, v_2, …, v_n. The cutoff for the given time period is set at two years, resulting in the following n-1 data points: v1 compared to v2 and v3 (taking the maximal difference), v2 compared to v3 and v4 (taking the maximal difference), etc. Further models using this patient's data are also created, set at different time periods (three years, four years, and so on).

There were two reasons for choosing the maximal difference as the label. Firstly, the main clinical value is in predicting whether a patient will develop DR, not in which eye it will be. Secondly, correlation found between the maximal right and left eye differences was relatively high (0.5), which indicates that difference between the eyes may well be incidental.

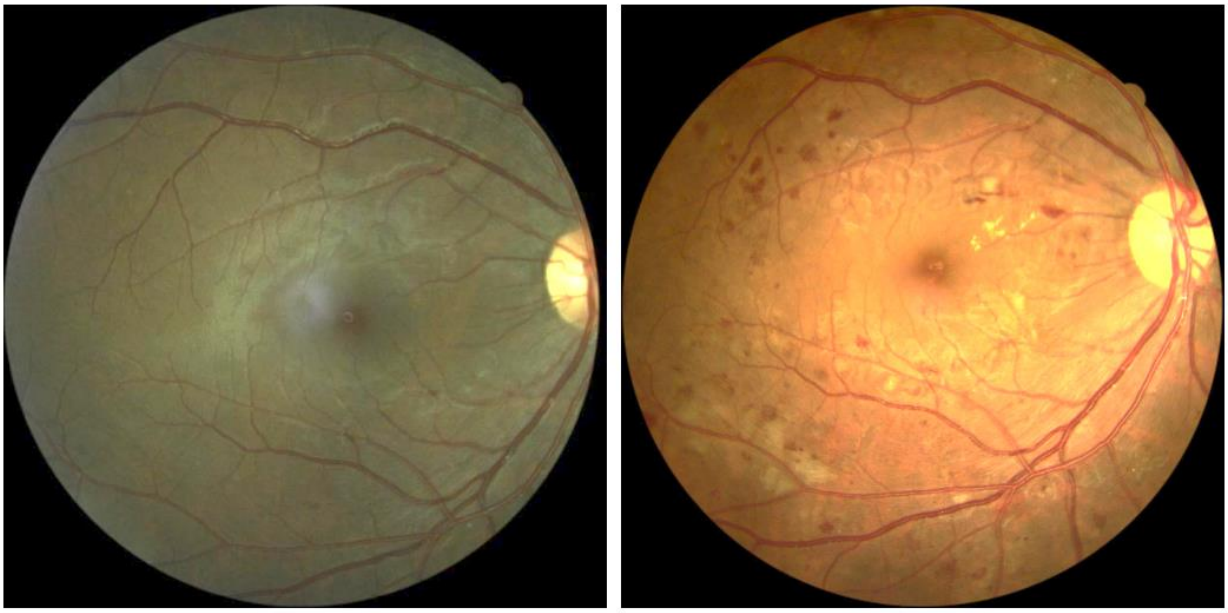


Figure 1: Left: the right eye at the first timepoint, rated healthy by a human expert. Right: the same eye one year later, with severe DR. The model score for both healthy eyes was in the top 2% of severity, implying more than fivefold the baseline risk of developing more than mild DR in the following two years. One year later, the patient was diagnosed with severe and proliferative DR in the right and left eyes respectively.

2.2 Algorithm development

To evaluate the models' performance, a random 10% of the patients were designated as the validation set and not used for the training of models. Given that this same 10% were used as the validation set across tasks, this choice made the fair comparison of different models and timeframes easier.

In order to train the model, all the datapoints representing a progression were included, and a subset of the negative datapoints were included at a ratio of 2:1.

Models were trained on four different tasks:

- Progression amongst DR patients (mild to more-than-mild)
- Prediction of DR development (normal to any DR)
- Prediction of clinically significant DR (from non-referrable to referrable DR)
- General progression (progress): any change for the worse in the DR condition

The hyper-parameters for the model training were chosen beforehand, and not changed, to prevent over-fitting.

2.3 Risk factor predictive value

For 80% of the patients in the dataset, HbA1c level was recorded by the clinic, and disease duration was recorded for 98% of patients. HbA1c level and disease duration were treated as risk level scores. AUC was then calculated in order to rate predictive value for each of the scores, which were then compared to the predictive value of the model on each task (see table [[whatever]]).

3. Results

3.1 Transitional Retinopathy Results

The calculated/resulting baseline transition odds between different DR levels are displayed in **Error! Reference source not found.** (for a more detailed table see appendix B). This is observational data, as regression- and progression-related factors are unknown; regression may have been caused by clinical intervention, and progression may be understated due to patients who experienced vision loss and therefore did not return for subsequent visits.

Initial DR level		0	1	2	3	4
Year 1	Regression	-	0.413	0.328	0.35	0.119
No Change		0.921	0.38	0.613	0.567	0.881
Progression		0.079	0.208	0.059	0.084	-
Year 2	Regression	-	0.383	0.307	0.461	0.159
No Change		0.895	0.328	0.596	0.34	0.841
Progression		0.105	0.288	0.097	0.198	-
Year 3	Regression	-	0.423	0.285	0.424	0.105
No Change		0.872	0.303	0.624	0.364	0.895
Progression		0.128	0.274	0.091	0.212	-

Table 2: Disease progression was calculated between any given visit and the visit immediately following. "Regression" is defined as the patient's recorded DR level being lower on the second visit, "No change" is defined as recorded DR levels being the same between visits, and "Progression" is defined as DR levels being higher.

3.2 Prediction Results

The model's performance in determining the risk of mild DR becoming more-than-mild DR is comparable to risk factor-based prediction (area under the receiver operating curve (AUC) 0.65 vs. 0.66, respectively). For the other tasks more images were available, (see appendix J in supplementary) and performance improved significantly. The results improved still by using multiple images per patient and averaging the resulting score. The model scored best on the task of prediction of clinically significant DR, with the aggregated score resulting in AUC 0.81 (CI 95% 0.77-0.84) (see **Error! Reference source not found.**. Additional timeframes and ROC curves available in appendix G and H). As the HbA1c levels and disease duration was not available for all patients, model scores were also calculated for the subsets of patients who had those scores, resulting in effectively the same scores (see appendix M for exact numbers). The correlation between our model's score and HbA1c levels was 0.12 (CI 95% 0.06-0.21). Correlation with disease duration was 0.21 (CI 95% 0.22-0.30).

	Image Prediction (by algorithm)	Patient Prediction (by algorithm)	HbA1c-based prediction	Disease Duration-based prediction

Mild DR to mtmDR	0.63 (0.59, 0.67)	**0.65 (0.57, 0.73)**	**0.66 (0.62, 0.71)**	0.50 (0.46, 0.54)
No DR to mtmDR	0.67 (0.65, 0.68)	**0.71 (0.68, 0.74)**	0.65 (0.63, 0.67)	0.60 (0.59, 0.62)
mtmDR- to mtmDR+	0.75 (0.73, 0.76)	**0.81 (0.77, 0.84)**	0.70 (0.68, 0.72)	0.63 (0.62, 0.65)
Any DR Progression	0.70 (0.69, 0.72)	**0.75 (0.72, 0.78)**	0.67 (0.66, 0.69)	0.61 (0.60, 0.62)

Table 3: The models' score, in AUC, in predicting DR within two years. In the parentheses 95% CI. See supplemental for additional timeframes.

To further analyze the model's prediction value in that task, the empirical risk as a function of the model's score was investigated (see **Error! Reference source not found.**). When the model was trained to predict the transition to more-than-mild DR, the top 5% patients highest scored by the model were at 54% risk of getting DR, while the baseline odds in the validation set were 10% - almost a 5-fold increase.

In order to analyze the relation between the model's assigned scores and the severity of developed DR, the

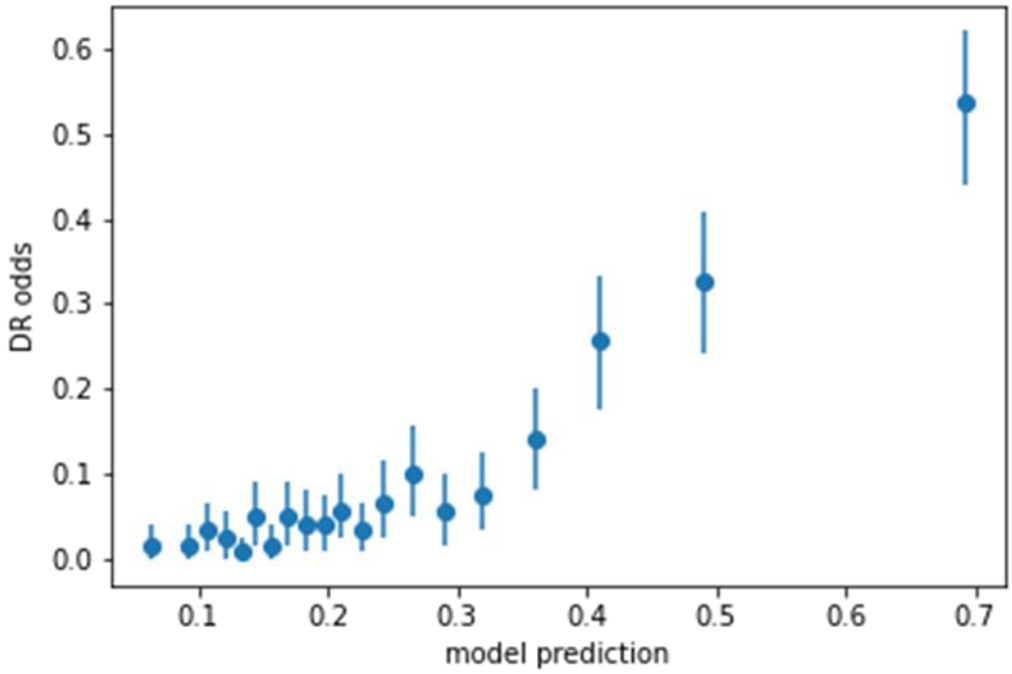


Figure 2: the odds of the patient from a representative sample being diagnosed with clinically significant DR in two years, as a function of the model's score. Each dot represents 5% of the patients. Additional figures in appendix K

scores assigned by the model were averaged and compared across different subgroups; normal patients who were diagnosed with DR up to two years after initial images were taken were sorted into subgroups based on DR severity. The groups of mild, moderate, and severe DR were score averaged at 0.36, 0.43, and 0.46 respectively (see **Error! Reference source not found.Error! Reference source not found.**).

Additionally, the model's predictive effectiveness regarding the more rapid development of DR was analyzed. Assigned scores were averaged and compared across subgroups organized by time elapsed between initial imaging and diagnosis: one, two, three, and four or more years. It was hypothesized that

the healthier the eye appeared the more time would elapse between initial imaging and diagnosis. As expected, the model's score declined in congruence with time passed before diagnosis (see Figure 4).

Figure 4: The model's mean score as a function of how many years after the visit DR was diagnosed. As expected, model average score

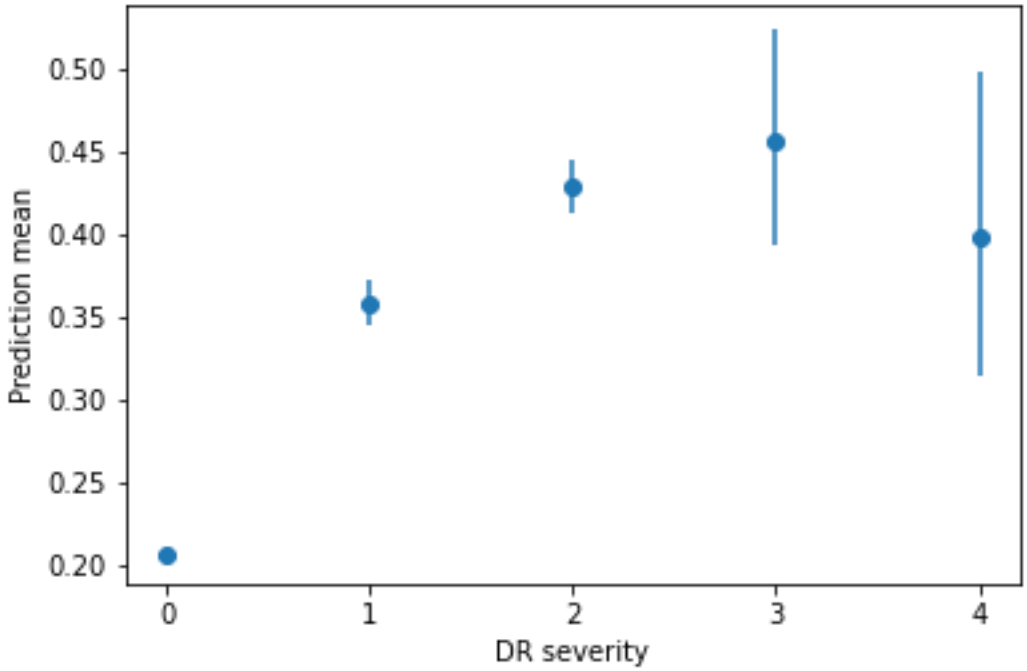
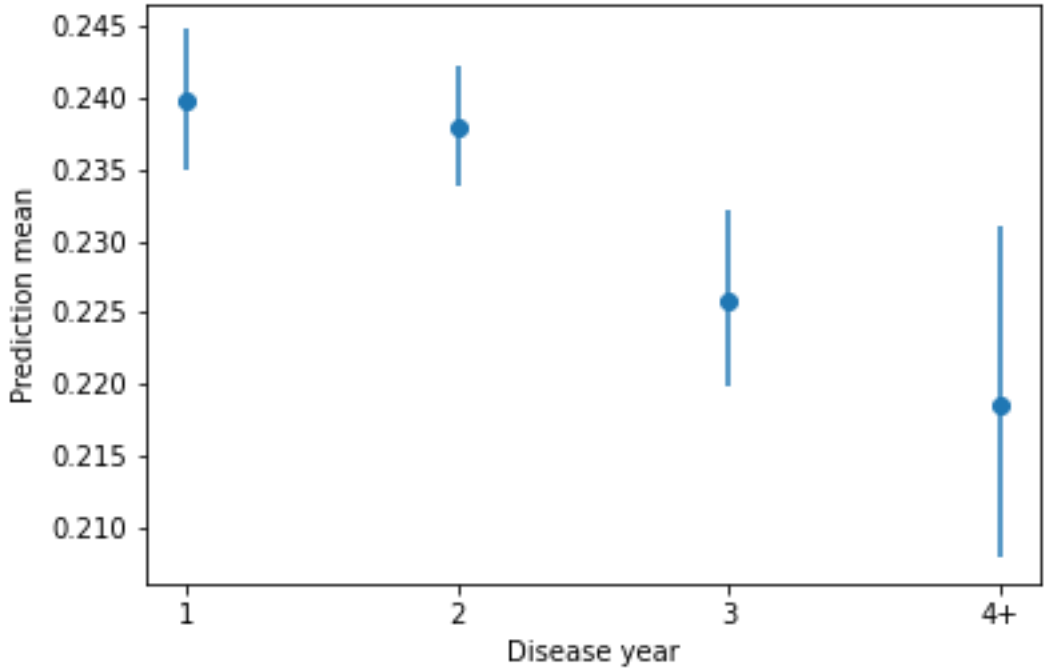


Figure 3: The model's score as a function of the final diagnosed severity of the DR. The model is more confident of the future occurrence of DR if it's more severe.



decreases in accordance with time elapsed between first visit and diagnosis.

4. Discussion

The aim of this research was to develop a method of predicting the chances of future development of DR, before detection methods can even be applied. Given the extremely high prevalence of DR,[2] the difficulty in permanently reversing retinal damage,[24] and current patient non-compliance,[8–11] prevention and not only treatment of DR is crucial to health outcomes. As such, prediction of individual DR risk may become a key element. Currently, to the best of our knowledge, the only methods of doing so are risk factor based and are of limited predictive clinical utility.[15,16]

The current research was conducted using convolutional neural networks, a standard state of the art computer vision algorithm. Such algorithms have been reliably incorporated in multiple medical fields, such as ophthalmology,[25] radiology,[26] endocrinology,[27] and others.[28] The algorithms presented in this work, which are easily implemented and display excellent performance, may carry widespread implications related to better DR prediction. For instance, given the knowledge that some patients are at very low risk of developing DR, screening may be able to be feasibly reduced according to individual risk levels, reducing strain placed on both patients and medical staff. Furthermore, in high-risk cases which have not yet manifested, patients may be forewarned of impending risk, increasing chances of mitigation and prevention through diabetes management.

A lack of patient education regarding the risks of diabetes, and DR specifically, has been cited as a contributing factor in patient non-compliance.[8] Furthermore, patients may not attend screenings due to belief that they do not require retinal examinations or treatment as their vision is too good, or their diabetes is too mild to be relevant. The ability to concretely discuss personal risk levels with the patient may do much to mitigate these beliefs, contributing to higher compliance. Improved patient compliance in terms of DR may also improve compliance in terms of general diabetes management, bettering patient outcomes across the board.

One limitation of the current research is that the transition odds are observatory, rather than experimental. As such, there is the possibility of an under-statement of risk, given that blinded people likely did not continue to return for checkups. Odds of regression may similarly be over-stated, as regression factors are unknown and regression may have been caused by surgical or medical intervention.

Recommendations for future research include studies on how to incorporate the model into usual diabetes standards of care, in order to better mitigate and prevent DR. As the model's score was not strongly correlated to risk factor score, there may be added value in including metadata on levels of previously recognized risk factors among patients in order to improve predictive value. Additionally, there is great value in examining whether use of this algorithm does, in fact, improve patient compliance. This model may also contribute to future research of DR risk factors and prevention, as at-risk patients with previously unknown risk factors may become recognizable, allowing for a more holistic understanding of contributing influences, both biological and behavioral.

5. Bibliography

1 Cheung N, Mitchell P, Wong TY. Diabetic retinopathy. *Lancet* 2010; **376**: 124–36.

2 Kertes PJ, Johnson TM. Evidence-based eye care. Lippincott Williams & Wilkins, 2007.

3 Tapp RJ, Shaw JE, Harper CA, *et al.* The prevalence of and factors associated with diabetic retinopathy in the Australian population. *Diabetes care* 2003; **26**: 1731–7.

4 American Diabetes Association Professional Practice Committee, Draznin B, Aroda VR, *et al.* Retinopathy, Neuropathy, and Foot Care: Standards of Medical Care in Diabetes-2022. *Diabetes Care* 2022; **45**: S185–94.

5 World Health Organization. Regional Office for Europe. Diabetic retinopathy screening: a short guide: increase effectiveness, maximize benefits and minimize harm. World Health Organization. Regional Office for Europe, 2020 https://apps.who.int/iris/handle/10665/336660 (accessed Feb 27, 2022).

6 Rohan TE, Frost CD, Wald NJ. Prevention of blindness by screening for diabetic retinopathy: a quantitative assessment. *BMJ* 1989; **299**: 1198–201.

7 Stefánsson E, Bek T, Porta M, Larsen N, Kristinsson JK, Agardh E. Screening and prevention of diabetic blindness. *Acta Ophthalmol Scand* 2000; **78**: 374–85.

8 Lewis K. Improving patient compliance with diabetic retinopathy screening and treatment. *Community Eye Health* 2015; **28**: 68–9.

9 National Institute for Health and Care Excellence (NICE). Type 2 diabetes in adults: management. NICE https://www.nice.org.uk/guidance/ng28 (accessed Feb 27, 2022).

10 Paz SH, Varma R, Klein R, Wu J, Azen SP, Los Angeles Latino Eye Study Group. Noncompliance with vision care guidelines in Latinos with type 2 diabetes mellitus: the Los Angeles Latino Eye Study. *Ophthalmology* 2006; **113**: 1372–7.

11 Saadine JB, Fong DS, Yao J. Factors associated with follow-up eye examinations among persons with diabetes. *Retina* 2008; **28**: 195–200.

12 Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *npj Digital Med* 2018; **1**: 1–8.

13 AEYE. AEYE Health Reports Pivotal Clinical Trial Results of its AI Algorithm for the Autonomous Screening and Detection of More-Than-Mild Diabetic Retinopathy. https://www.prnewswire.com/news-releases/aeye-health-reports-pivotal-clinical-trial-results-of-its-ai-algorithm-for-the-autonomous-screening-and-detection-of-more-than-mild-diabetic-retinopathy-301476299.html (accessed Feb 28, 2022).

14 Dvey-Aharon Z, Huhtinen P. Screening for Diabetic Retinopathy in Endocrinology Clinics by Using Handheld Cameras and Applying Artificial Intelligence Algorithms. *J Endocr Soc* 2021; **5**: A419–20.

15 Magliah SF, Bardisi W, Al Attah M, Khorsheed MM. The prevalence and risk factors of diabetic retinopathy in selected primary care centers during the 3-year screening intervals. *J Family Med Prim Care* 2018; **7**: 975–81.

16 Sheu S-J, Liu N-C, Ger L-P, *et al.* High HbA1c level was the most important factor associated with prevalence of diabetic retinopathy in Taiwanese type II diabetic patients with a fixed duration. *Graefes Arch Clin Exp Ophthalmol* 2013; **251**: 2087–92.

17 Sabanayagam C, Xu D, Ting DSW, *et al.* A deep learning algorithm to detect chronic kidney disease from retinal photographs in community-based populations. *The Lancet Digital Health* 2020; **2**: e295–302.

18 Mitani A, Hammel N, Liu Y. Retinal detection of kidney disease and diabetes. *Nat Biomed Eng* 2021; **5**: 487–9.

19 Poplin R, Varadarajan AV, Blumer K, *et al.* Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng* 2018; **2**: 158–64.

20 Haneda S, Yamashita H. International clinical diabetic retinopathy disease severity scale. *Nihon rinsho Japanese journal of clinical medicine* 2010; **68**: 228–35.

21 Gulshan V, Peng L, Coram M, *et al.* Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* 2016; **316**: 2402–10.

22 Flaxel CJ, Adelman RA, Bailey ST, *et al.* Diabetic Retinopathy Preferred Practice Pattern®. *Ophthalmology* 2020; **127**: P66–145.

23 American Optometric Association. Eye Care of the Patient with Diabetes Mellitus. US, 2019 https://www.aoa.org/AOA/Documents/Practice%20Management/Clinical%20Guidelines/EBO%20Guidelines/Eye%20Care%20of%20the%20Patient%20with%20Diabetes%20Mellitus%2C%20Second%20Edition.pdf.

24 Wykoff CC, Eichenbaum DA, Roth DB, Hill L, Fung AE, Haskova Z. Ranibizumab Induces Regression of Diabetic Retinopathy in Most Patients at High Risk of Progression to Proliferative Diabetic Retinopathy. *Ophthalmology Retina* 2018; **2**: 997–1009.

25 Food and Drug Administration (FDA). DE NOVO CLASSIFICATION REQUEST FOR IDX-DR. Coralville, IA, 2018 https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN180001.pdf (accessed Feb 28, 2022).

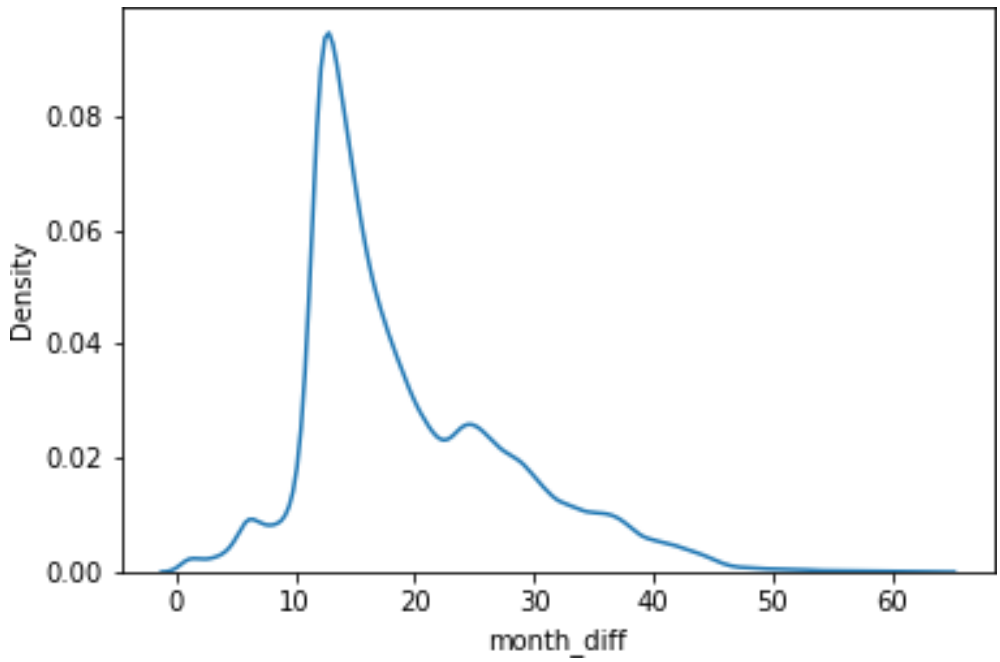
26 Food and Drug Administration (FDA). EVALUATION OF AUTOMATIC CLASS III DESIGNATION FOR OSTEODETECT. Coralville, IA, 2018 https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN180005.pdf (accessed Feb 28, 2022).

27 Food and Drug Administration (FDA). P160007 Approval Order. Coralville, IA, 2018 https://www.accessdata.fda.gov/cdrh_docs/pdf16/P160007A.pdf (accessed Feb 28, 2022).

28 Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *npj Digit Med* 2020; **3**: 1–8.

6. Appendices

6.1 A


0.08
0.06
0.04
0.02
0.00
Density
0
10
20
30
40
50
60
month_diff

6.2 B

DR progression, by initial stage. Regression includes recovery by surgery.

One year:

	0	1	2	3	4
-4	-	-	-	-	0.0042
-3	-	-	-	0.0083	0.0084
-2	-	-	0.199	0	0.0816
-1	-	0.4127	0.129	0.3412	0.0249
0	0.9209	0.3798	0.6129	0.5667	0.8809
1	0.044	0.2045	0.0462	0.0838	-
2	0.0343	0.0026	0.0129	-	-
3	0.0004	0.0004	-	-	-
4	0.0004	-	-	-	-

Two years:

	0	1	2	3	4
-4	-	-	-	-	0.0111
-3	-	-	-	0.0175	0.0148
-2	-	-	0.1683	0.0175	0.0888
-1	-	0.3834	0.1391	0.4261	0.0444
0	0.8951	0.3284	0.596	0.3405	0.8409
1	0.0576	0.2784	0.0667	0.1984	-
2	0.0461	0.0058	0.03	-	-
3	0.0008	0.0039	-	-	-
4	0.0003	-	-	-	-

Three years:

	0	1	2	3	4

-4	-	-	-	-	0
-3	-	-	-	0.0909	0
-2	-	-	0.1725	0.0909	0.1053
-1	-	0.4229	0.1128	0.2424	0
0	0.8719	0.303	0.6241	0.3636	0.8947
1	0.0746	0.2622	0.042	0.2121	-
2	0.0504	0.0119	0.0485	-	-
3	0.0023	0	-	-	-
4	0.0009	-	-	-	-

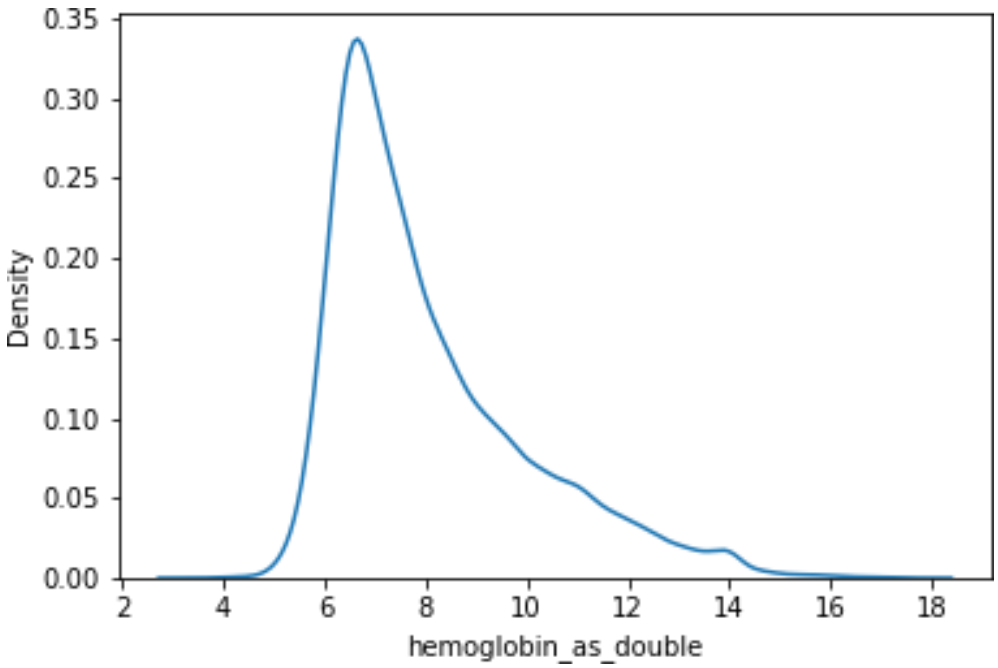
6.3 C

The images in the original dataset were taken using several cameras. The exact division is shown in the following table:

Camera Name	Number of Images
Cannon CR2	263,815
Centervue DRS	145,257
Crystalvue	8,092
Topcon NW400	367,854

6.4 D

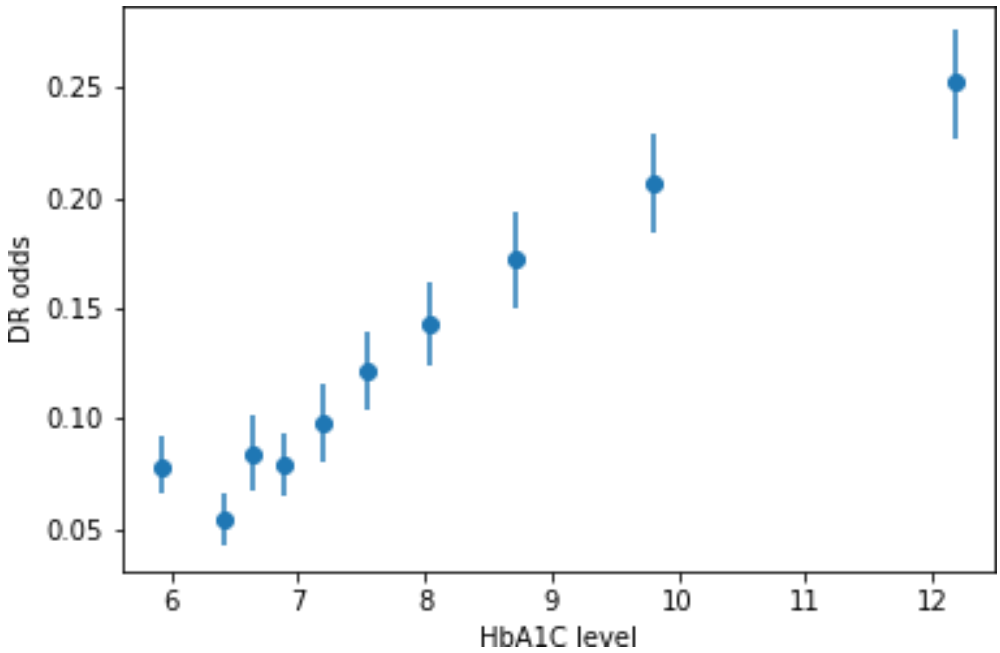
HbA1c distribution:



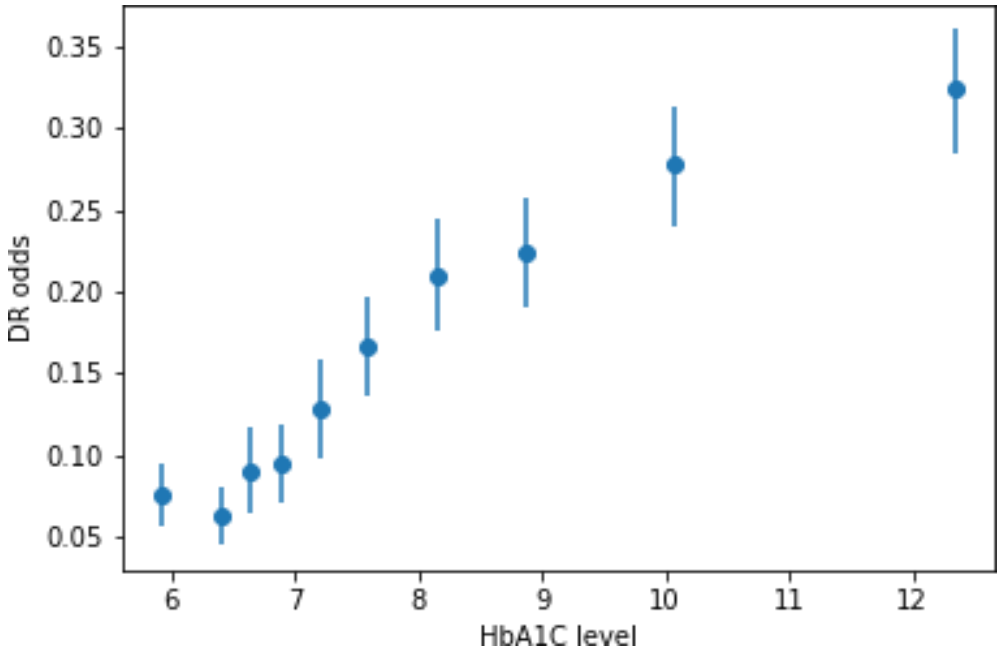
There were 20 patients with HbA1C values ranging from 56.2 to 1132 who were excluded from this figure

6.5 E

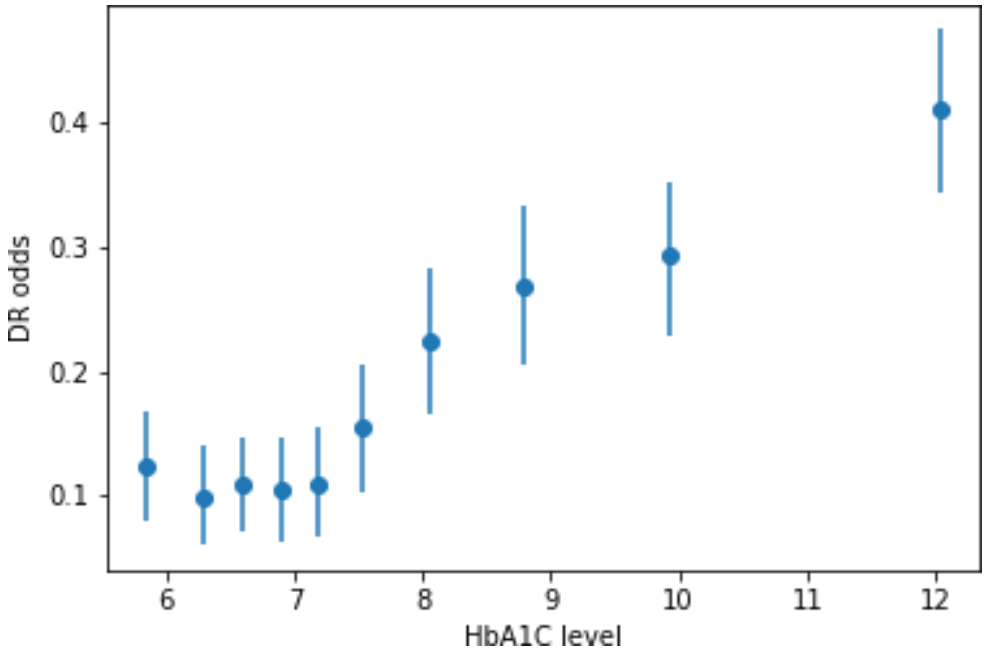
DR odds as a function of hemoglobin, first year:



DR odds as a function of HbA1c, second year:



DR odds as a function of HbA1c, third year:


DR odds
0.4
0.3
0.2
0.1
6
7
8
9
10
11
12
HbA1C level

6.6 F

Disease progression was calculated between any given visit and the visit immediately following. “Regression” is defined as the patient’s recorded DR level being lower on the second visit, “No change” is defined as recorded DR levels being the same between visits, and “Progression” is defined as DR levels being higher.

First year:

	specific image	specific patient
mild_to_severe	0.576 (0.515, 0.634)	0.618 (0.495, 0.731)
healthy_to_severe	0.641 (0.621, 0.661)	0.677 (0.626, 0.722)
to_severe	0.734 (0.710, 0.756)	0.789 (0.738, 0.831)
progress	0.670 (0.652, 0.688)	0.711 (0.666, 0.750)

Second year:

	specific image	specific patient
mild_to_severe	0.630 (0.590, 0.670)	0.653 (0.572, 0.728)
healthy_to_severe	0.667 (0.652, 0.681)	0.708 (0.677, 0.738)
to_severe	0.746 (0.729, 0.761)	0.807 (0.773, 0.837)
progress	0.703 (0.690, 0.716)	0.750 (0.722, 0.777)

Three years:

	specific image	specific patient
mild_to_severe	0.611 (0.572, 0.648)	0.635 (0.559, 0.705)
healthy_to_severe	0.667 (0.654, 0.680)	0.712 (0.683, 0.739)
to_severe	0.725 (0.711, 0.740)	0.779 (0.749, 0.807)
progress	0.710 (0.699, 0.721)	0.753 (0.727, 0.776)

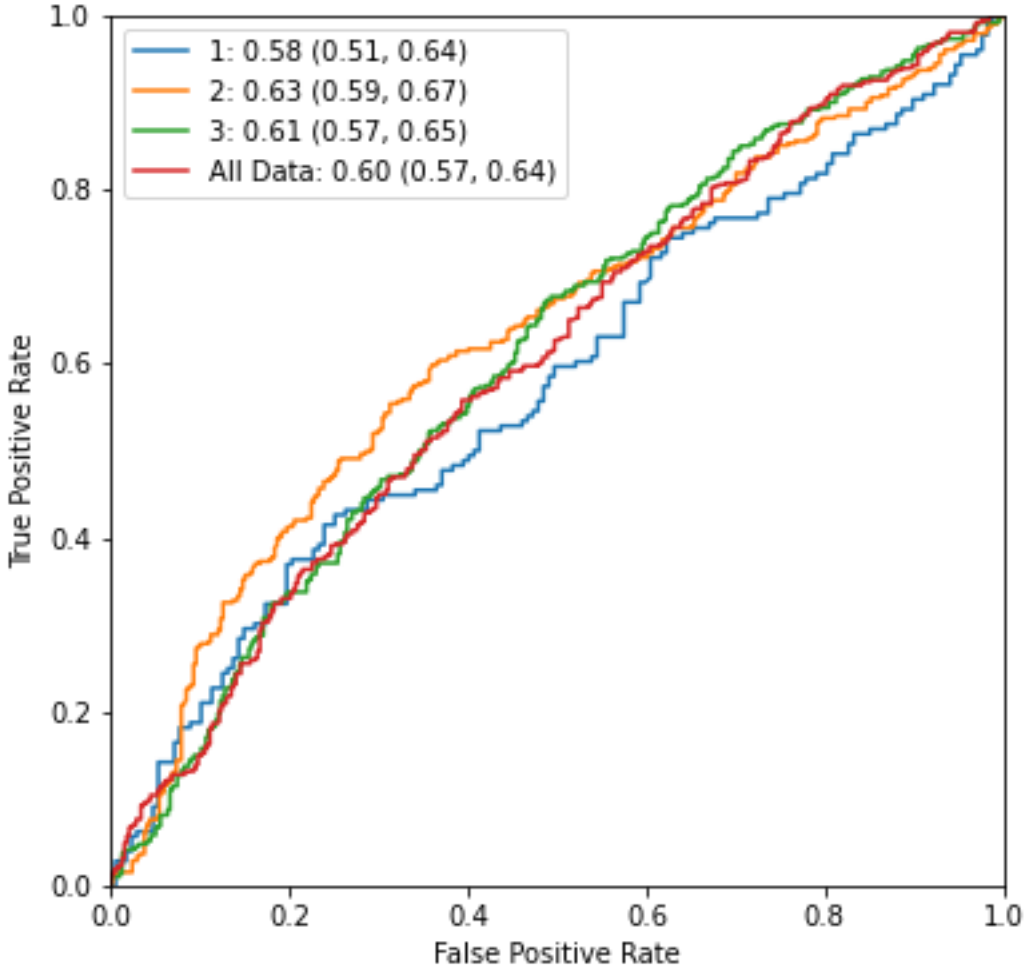
All images:

	specific image	specific patient
mild_to_severe	0.603 (0.566, 0.641)	0.629 (0.554, 0.697)
healthy_to_severe	0.669 (0.656, 0.682)	0.711 (0.683, 0.737)
to_severe	0.729 (0.714, 0.743)	0.785 (0.754, 0.811)
progress	0.705 (0.694, 0.716)	0.754 (0.729, 0.777)

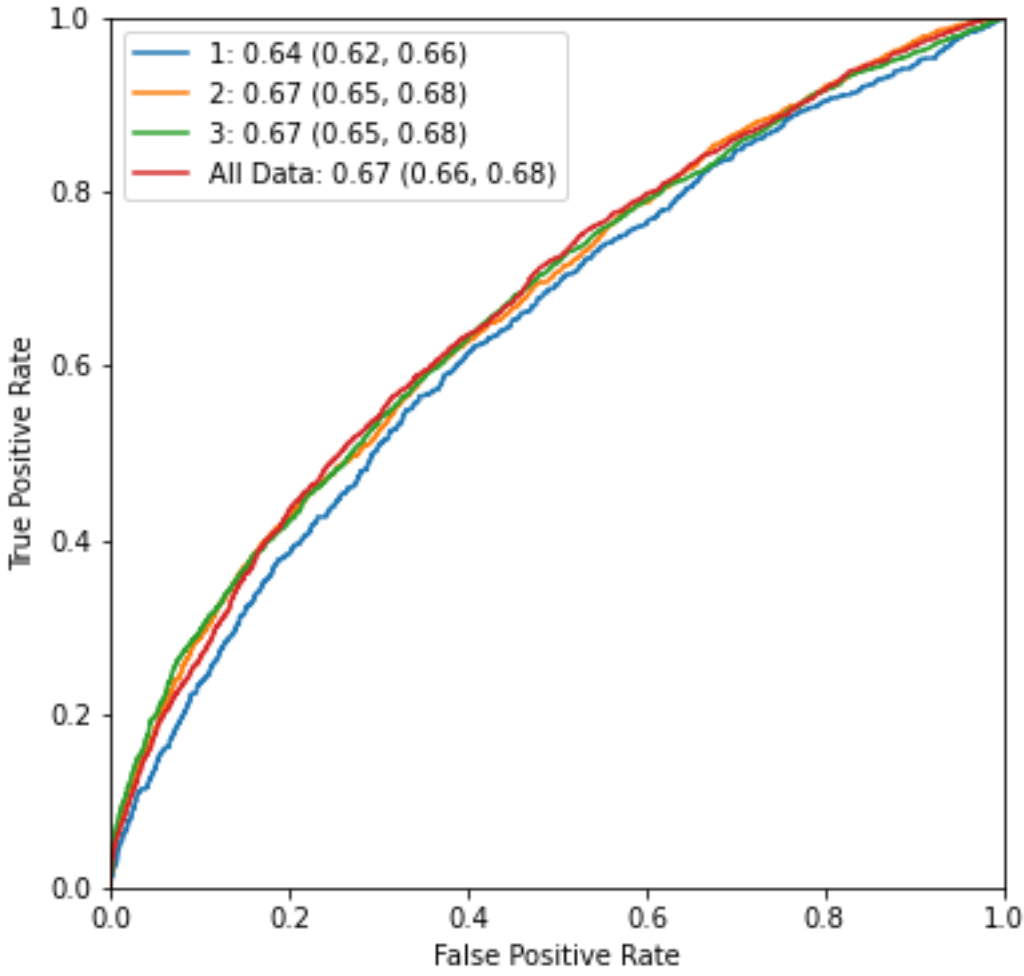
6.7 G

ROC curves of prediction by image.

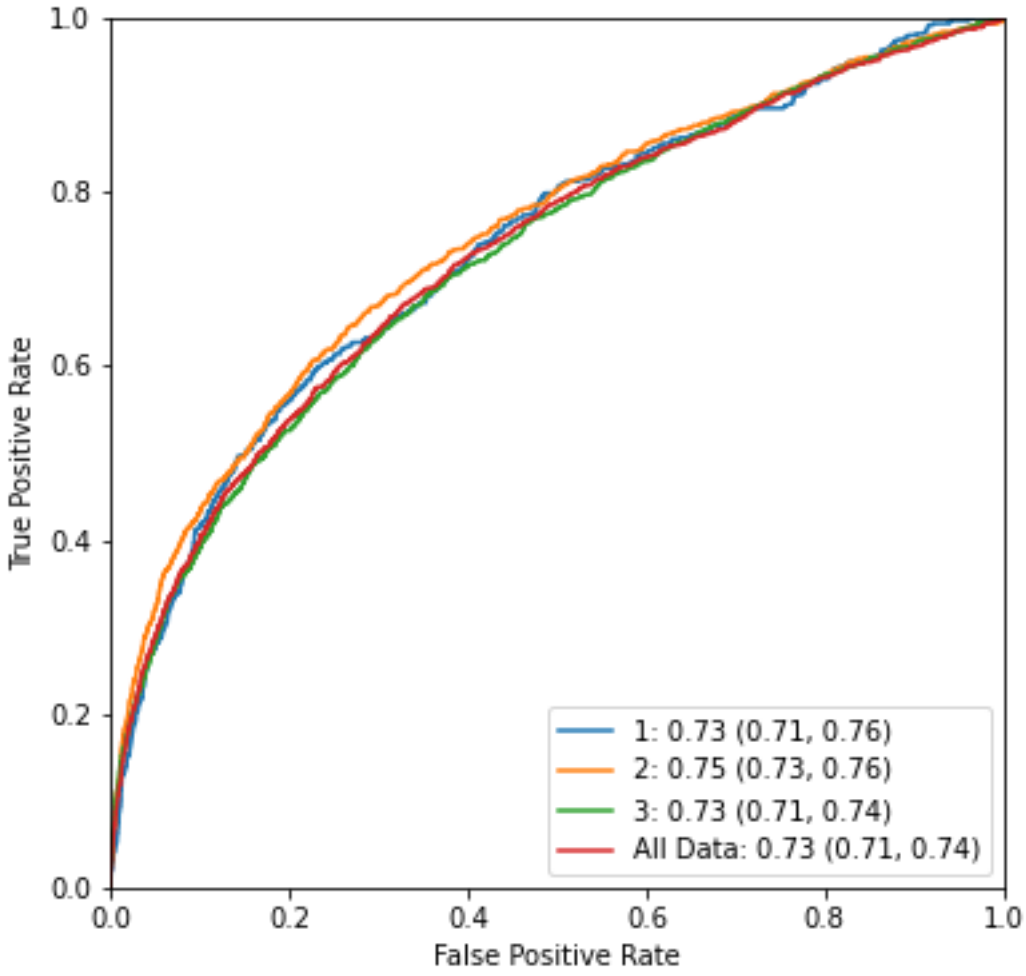
Mild to more than mild:



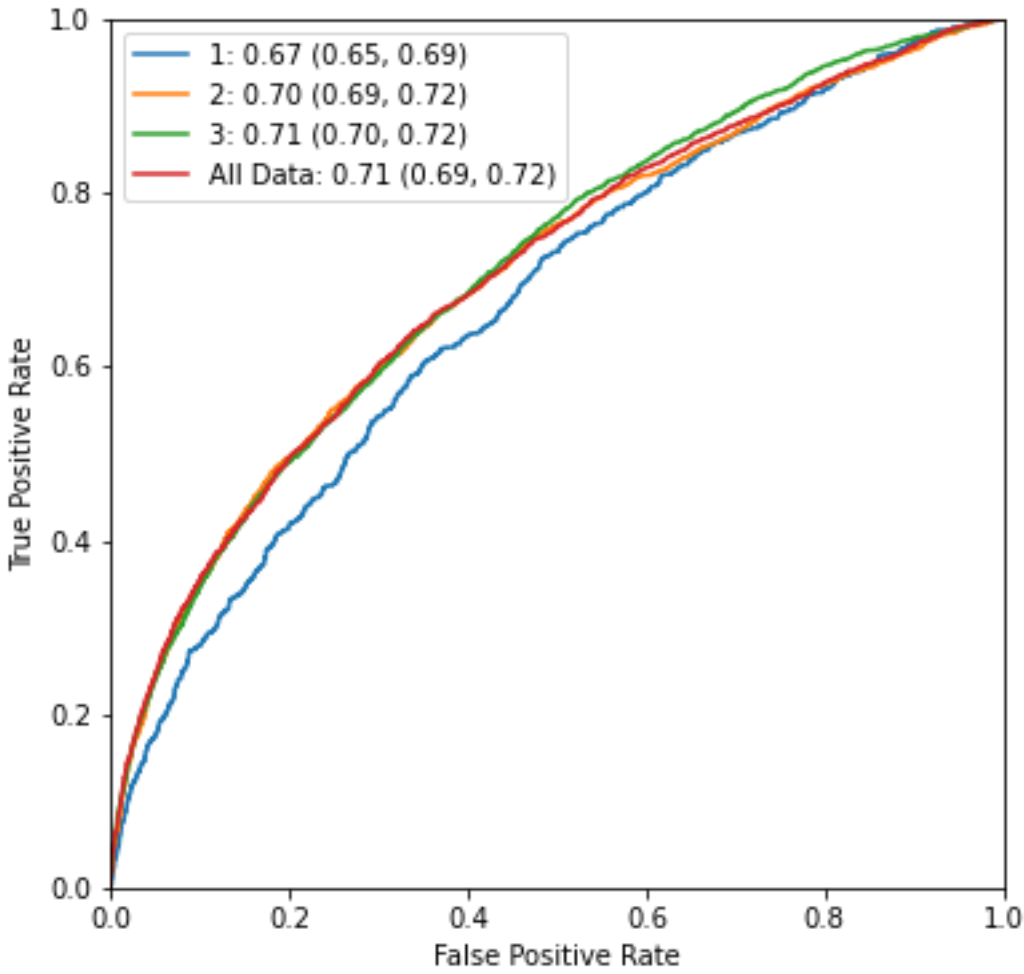
Normal to more than mild:



Non referrable to referrable:



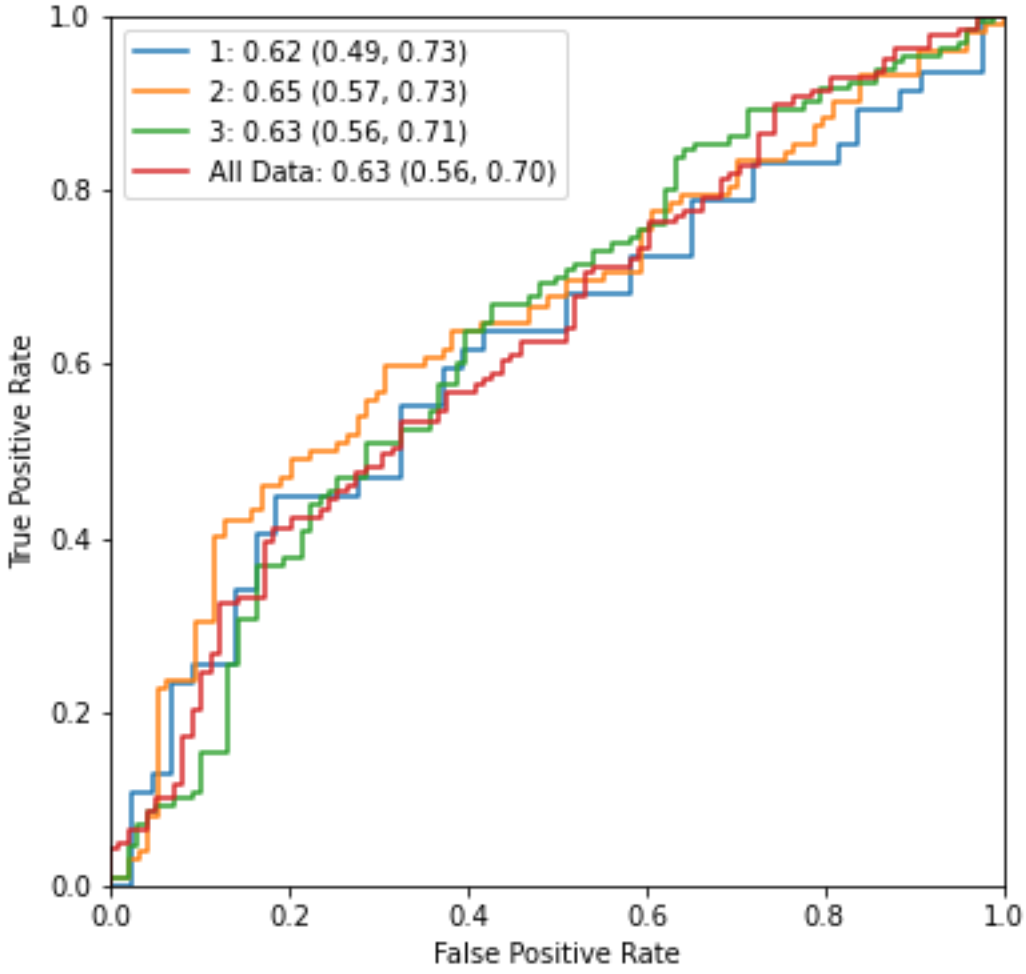
Any progression:



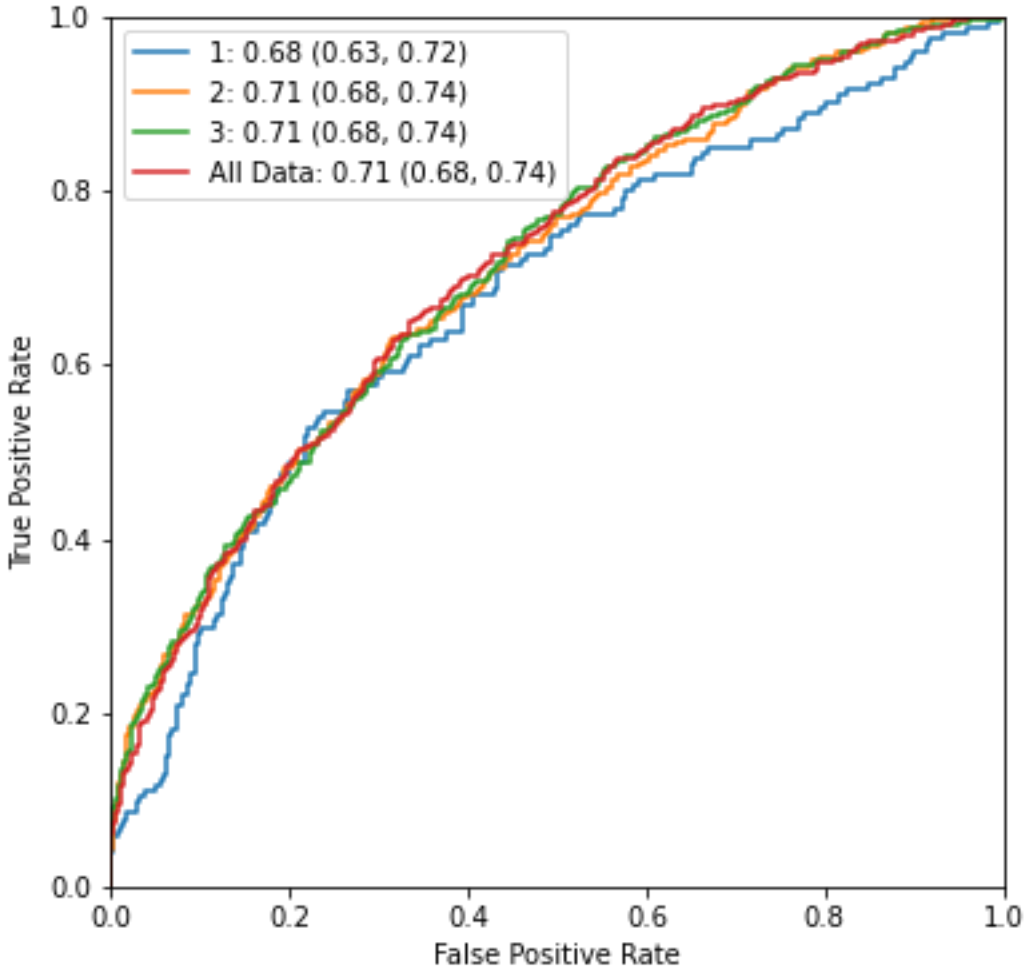
6.8 H

ROC curves of prediction by patient.

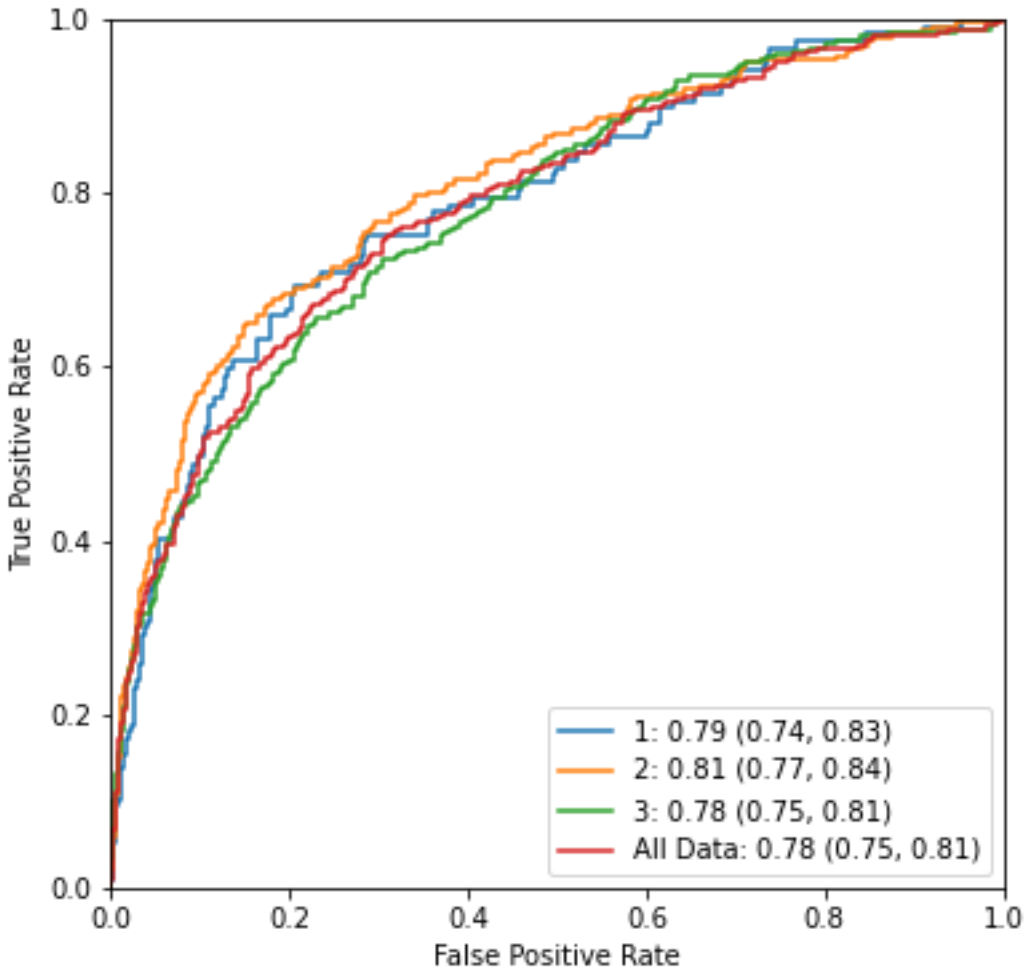
Mild to more than mild:



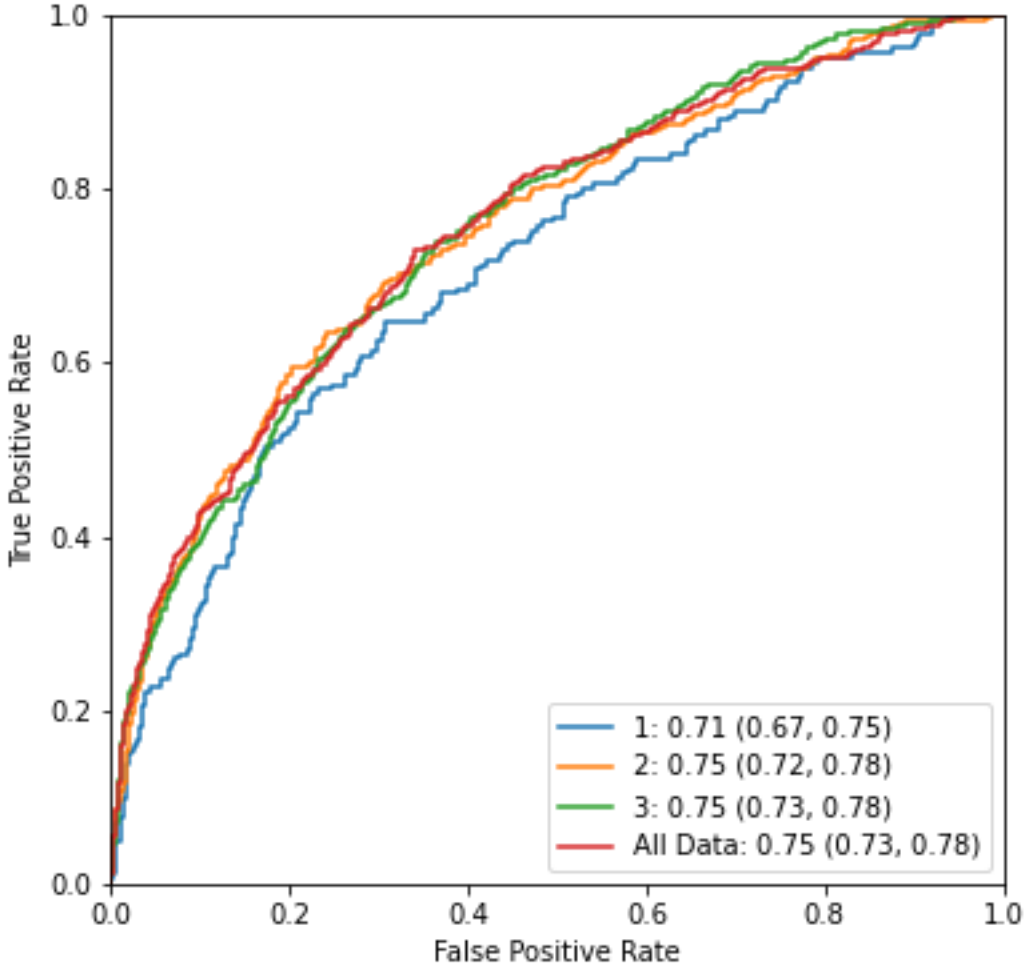
Normal to more than mild:



Non referrable to referrable:



Any progression:



6.9 I

Demographic characteristics of datasets with different timeframes:

First year:

	Training set	**Validation set**
Number of patients	8,947	968
Number of images	59,405	6,548
Age: mean, years (s.d.)	55.66 (10.52), n=8928	55.88 (10.17), n=966
Gender (% male)	0.37, n=8753	0.38, n=952
HbA1c: mean, % (s.d.)	7.87 (1.87), n=7159	7.93 (2.78), n=766
Disease duration: mean, years (s.d.)	7.62 (6.45), n=8640	7.35 (6.05), n=922
Ethnicity	62.6% Latin American, 11.6% ethnicity not specified, 8.7% African Descent, 6.4% Caucasian, 5.8% Asian, 3.3% Indian subcontinent origin, n=8740	62.8% Latin American, 11.1% ethnicity not specified, 8.8% African Descent, 5.9% Caucasian, 5.5% Asian, 3.9% Indian subcontinent origin, 1.4% Other, n=947

Second year:

	Training set	Validation set
Number of patients	17,161	1,896
Number of images	124,053	13,716
Age: mean, years (s.d.)	55.26 (10.65), n=17126	55.52 (10.41), n=1892
Gender (% male)	0.37, n=16783	0.36, n=1856
HbA1c: mean, % (s.d.)	7.96 (2.16), n=13845	7.95 (2.42), n=1515
Disease duration: mean, years (s.d.)	7.47 (6.43), n=16514	7.29 (6.27), n=1814
Ethnicity	62.8% Latin American, 11.4% ethnicity not specified, 8.7% African Descent, 7.2% Caucasian, 5.5% Asian, 2.7% Indian subcontinent origin, 1.1% Other, n=16753	61.5% Latin American, 10.8% ethnicity not specified, 9.1% African Descent, 7.8% Caucasian, 5.9% Asian, 3.2% Indian subcontinent origin, 1.2% Other, n=1856

Third year:

	Training set	Validation set
Number of patients	19,172	2,152
Number of images	138,333	15,447
Age: mean, years (s.d.)	55.19 (10.65), n=19136	55.35 (10.41), n=2148
Gender (% male)	0.37, n=18748	0.37, n=2109
HbA1c: mean, % (s.d.)	7.97 (2.26), n=15375	7.99 (2.76), n=1716
Disease duration: mean, years (s.d.)	7.46 (6.45), n=18448	7.20 (6.21), n=2061
Ethnicity	62.2% Latin American, 11.4% ethnicity not specified, 9.0% African Descent, 7.4% Caucasian, 5.5% Asian, 2.8% Indian subcontinent origin, 1.1% Other, n=18719	61.2% Latin American, 10.8% ethnicity not specified, 9.7% African Descent, 7.8% Caucasian, 5.8% Asian, 3.1% Indian subcontinent origin, 1.1% Other, n=2100

All data:

	Training set	Validation set
Number of patients	19,531	2,199
Number of images	140,614	15,749
Age: mean, years (s.d.)	55.15 (10.68), n=19495	55.34 (10.40), n=2195
Gender (% male)	0.37, n=19102	0.37, n=2155
HbA1c: mean, % (s.d.)	7.98 (2.26), n=15672	7.99 (2.75), n=1756
Disease duration: mean, years (s.d.)	7.46 (6.44), n=18803	7.21 (6.20), n=2108
Ethnicity	62.1% Latin American, 11.4% ethnicity not specified, 9.1% African Descent, 7.4% Caucasian, 5.4% Asian, 2.8% Indian subcontinent origin, 1.1% Other, n=19076	61.4% Latin American, 10.7% ethnicity not specified, 9.7% African Descent, 7.8% Caucasian, 5.6% Asian, 3.1% Indian subcontinent origin, 1.1% Other, n=2147

6.10 J

Table describing the sizes of datasets. First number is number of images in the development set, second is validation set.

	Mild DR to mtmDR	No DR to mtmDR	mtmDR- to mtmDR+	Any DR Progression
1	3415, 343	55990, 6205	59405, 6548	59405, 6548
2	7011, 745	117042, 12971	124053, 13716	124053, 13716
3	8121, 853	130212, 14594	138333, 15447	138333, 15447
4	8351, 885	132263, 14864	140614, 15749	140614, 15749

6.11 K

The empirical risk as a function of the model's score was investigated. The following four plots show the empirical risk as a function of model score per image for different tasks. The four following plots show the same per patient, by averaging the scores of all the images taken during one visit.

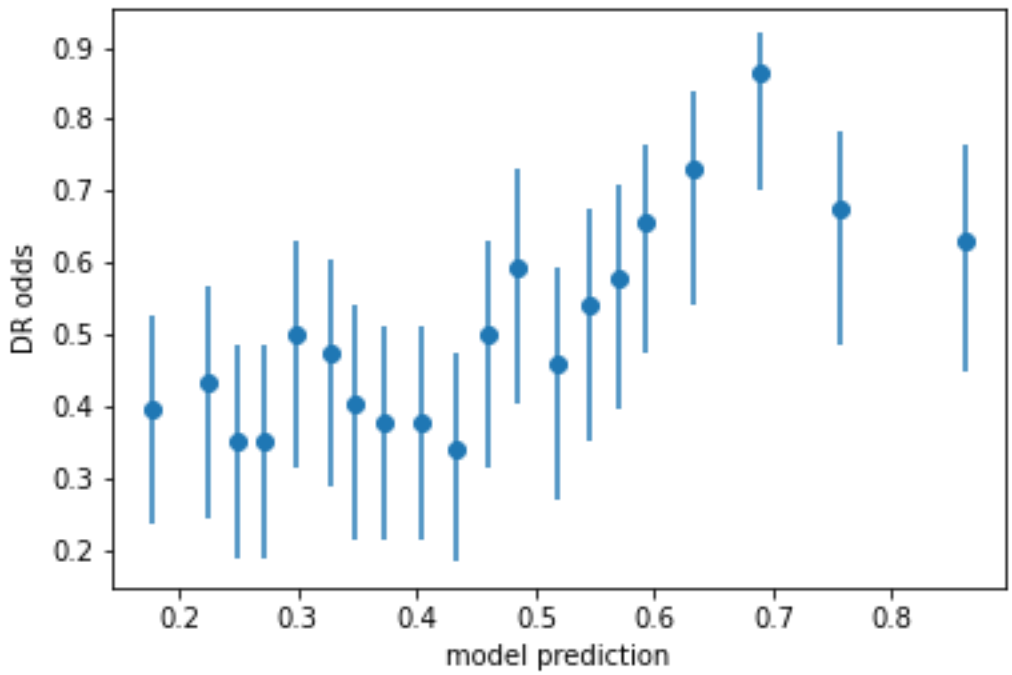


Figure 5: Empiric risk of mild DR to mtmDR as a function of model score per image

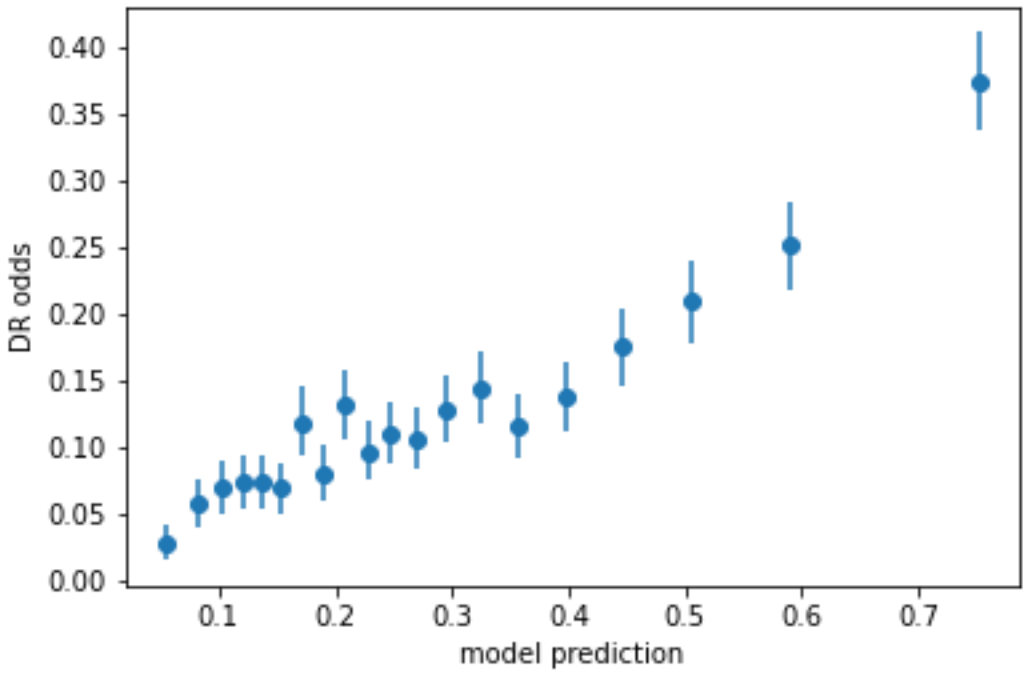


Figure 6: Empiric risk of no DR to mtmDR as a function of model score per image

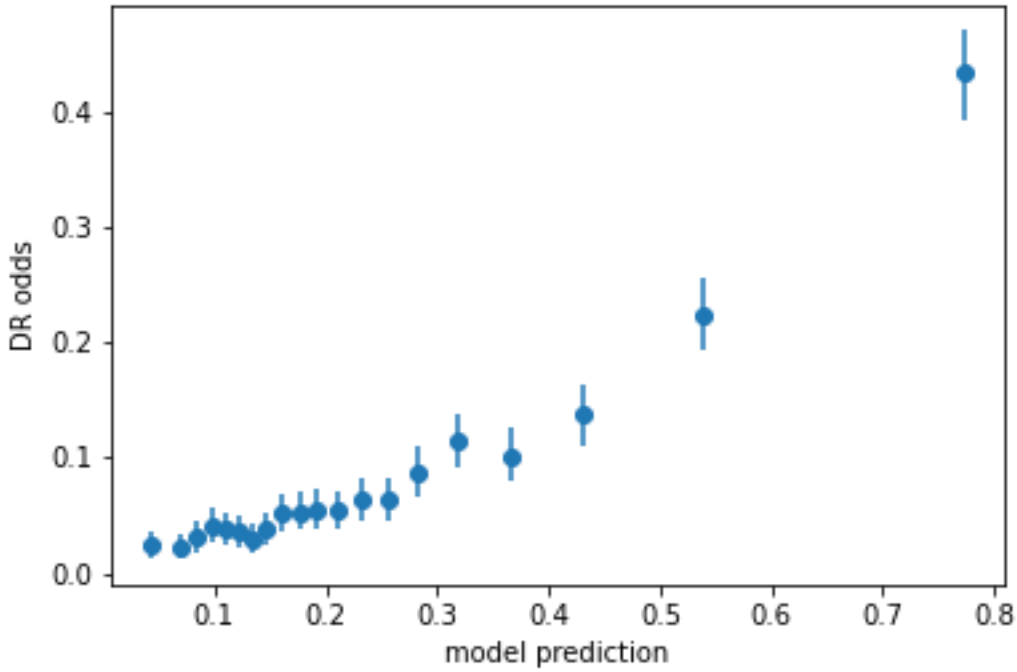


Figure 7: Empiric risk of mtmDR- to mtmDR+ as a function of model score per image

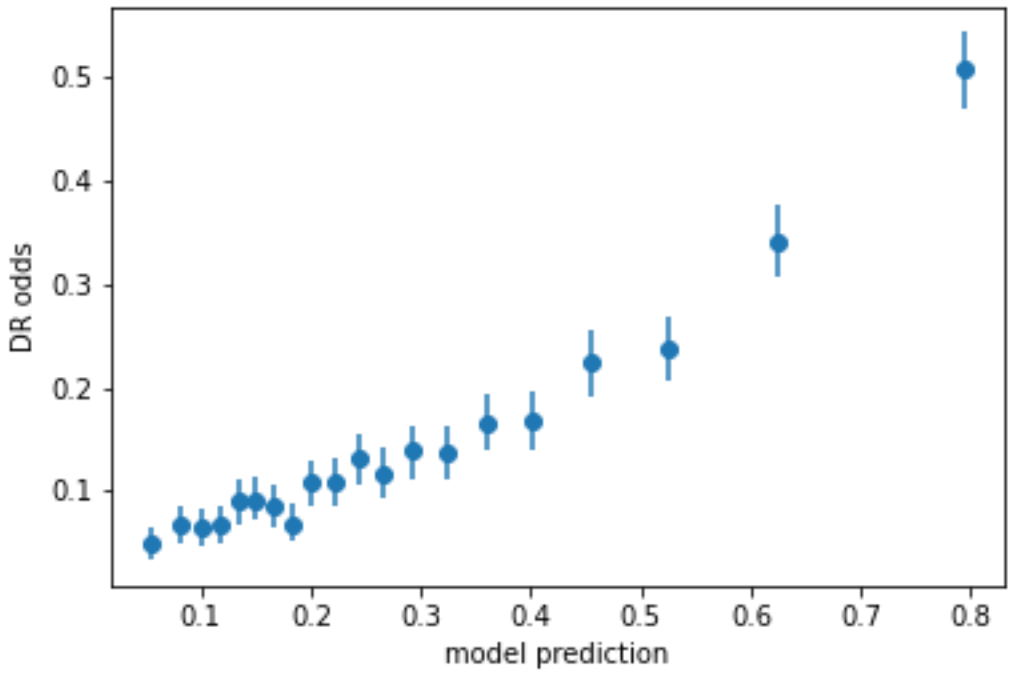


Figure 8: Empiric risk of any DR progression as a function of model score per image

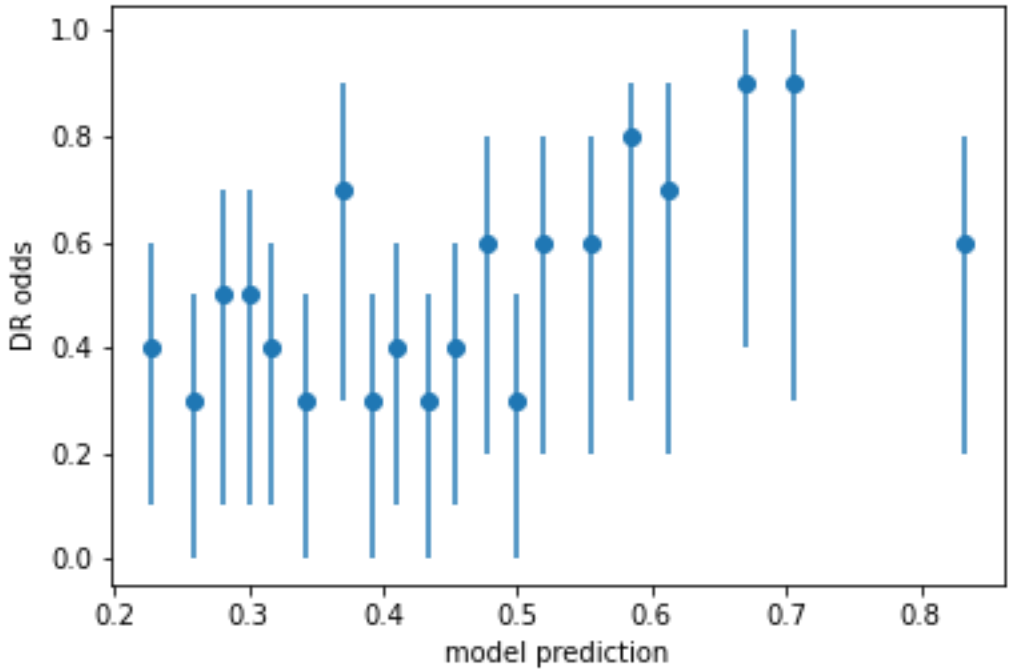


Figure 9: Empiric risk of mild DR to mtmDR as a function of model score per patient

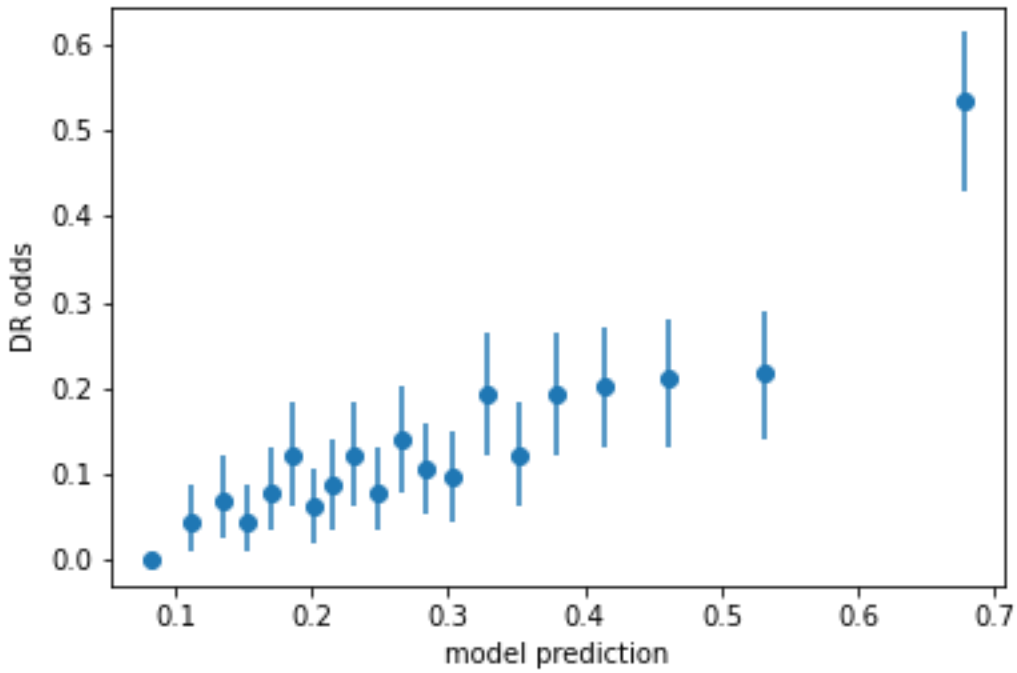


Figure 10: Empiric risk of no DR to mtmDR as a function of model score per patient

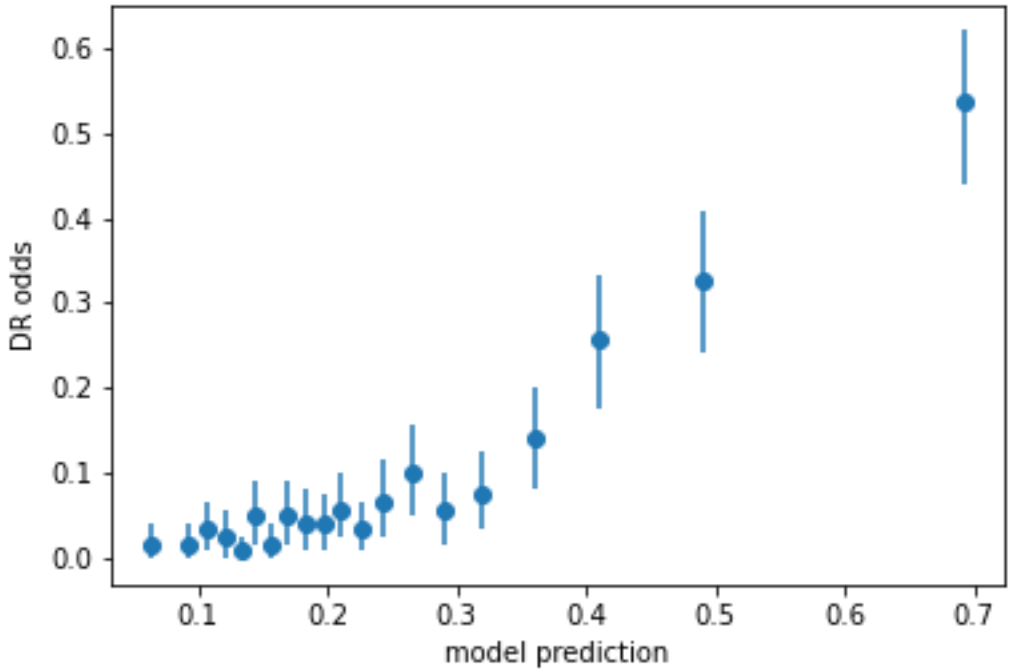


Figure 11: Empiric risk of mtmDR- to mtmDR+ as a function of model score per patient

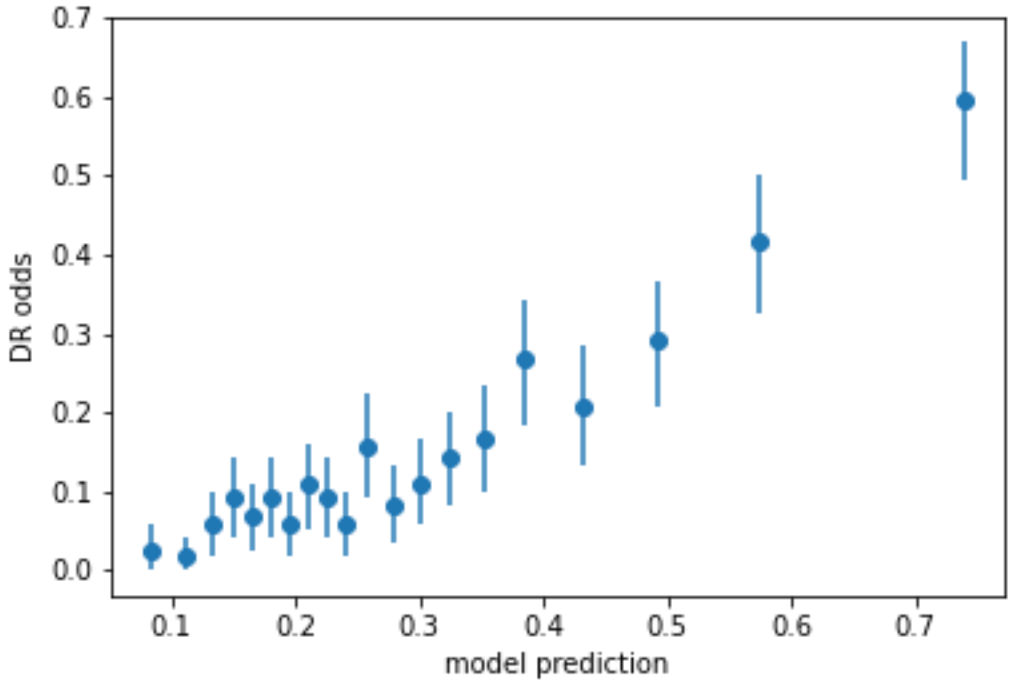
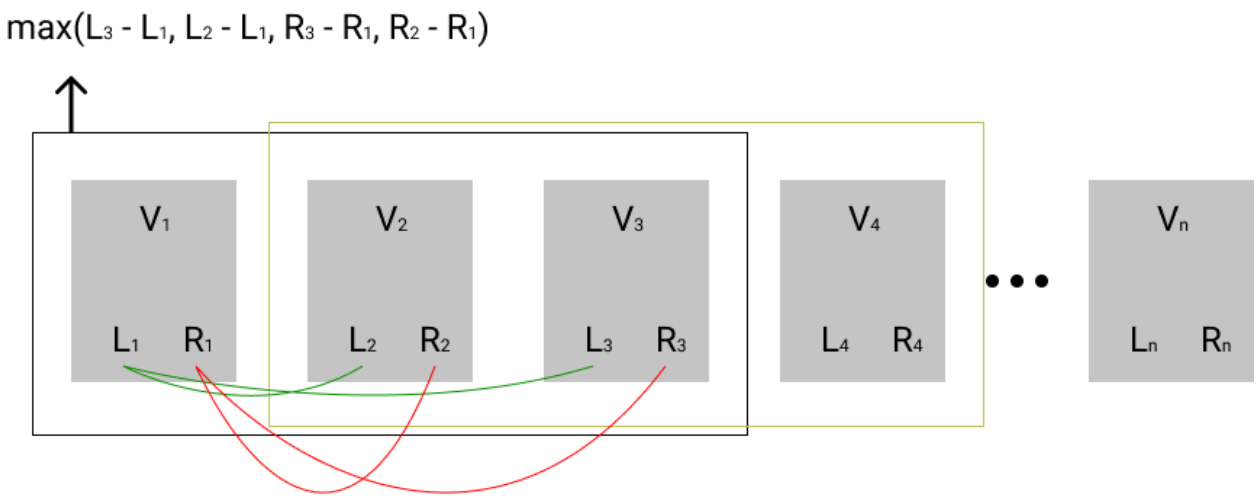


Figure 12: Empiric risk of any DR progression as a function of model score per patient

6.12 L

A figure describing our dataset creation process.



6.13 M

Table equivalent to table 1 in the results, describing model score only on the subset of patients who had HbA1c score or disease duration record.

	Image- only HbA1c	**Patient- only HbA1c**	**Image- only disease duration**	**Patient- only disease duration**
Mild DR to mtmDR	0.67 (0.63, 0.71)	0.69 (0.60, 0.77)	0.64 (0.59, 0.67)	0.66 (0.58, 0.73)
No DR to mtmDR	0.65 (0.64, 0.67)	0.69 (0.66, 0.73)	0.67 (0.65, 0.68)	0.71 (0.67, 0.74)
mtmDR- to mtmDR+	0.74 (0.72, 0.76)	0.80 (0.76, 0.84)	0.75 (0.73, 0.77)	0.81 (0.78, 0.84)
Any DR Progression	0.69 (0.68, 0.71)	0.75 (0.71, 0.78)	0.70 (0.69, 0.72)	0.75 (0.72, 0.78)