Predicting conversion to wet age related macular degeneration using deep learning

Authors: Jason Yim[1†], Reena Chopra[1,2†], Terry Spitz[3], Jim Winkens[3], Annette Obika[1], Christopher Kelly[3], Harry Askham[3], Marko Lukic[2], Josef Huemer[2], Katrin Fasler[2], Gabriella Moraes[2], Clemens Meyer[1], Marc Wilson[3], Jonathan Dixon[3], Cian Hughes[3], Geraint Rees[4], Peng T. Khaw[2], Alan Karthikesalingam[3], Dominic King[3], Demis Hassabis[1], Mustafa Suleyman[1], Trevor Back[1], Joseph R Ledsam[1‡]*, Pearse A. Keane[2‡]*, Jeffrey De Fauw[1‡]*

1 DeepMind, London, UK

2 NIHR Biomedical Research Centre at Moorfields Eye Hospital and UCL Institute of Ophthalmology, London, UK

3 Google Health, London, UK

4 University College London, London, UK

† Equal contribution, order determined randomly

‡ Equal contribution

* e-mail: pearse.keane1@nhs.net; defauw@google.com; jledsam@google.com

Abstract: Progression to exudative 'wet' age-related macular degeneration (exAMD) is a major cause of visual deterioration. In patients diagnosed with exAMD in one eye, we introduce an artificial intelligence (AI) system to predict progression to exAMD in the second eye. By combining models based on 3D optical coherence tomography images and corresponding automatic tissue maps, our system predicts conversion to exAMD within a clinically-actionable 6-month time window, achieving a per-volumetric-scan sensitivity of 80% at 55% specificity, and 34% sensitivity at 90% specificity. This level of performance corresponds to true positives in 78% and 41% individual eyes, and false positives in 56% and 17% individual eyes, at the high sensitivity and high specificity points respectively. Moreover, we show that automatic tissue segmentation can identify anatomical changes prior to conversion and high-risk subgroups. This AI system overcomes substantial interobserver variability in expert predictions, performing better than five out of six experts, and demonstrates the potential of using AI to predict disease progression.

Introduction

The application of artificial intelligence (AI) to disease classification has shown great promise towards increased utility and diagnostic accuracy for medical imaging[1-3]. Recent work has demonstrated further potential in risk stratification not previously thought possible[4-5]. There is significant potential for AI to improve our understanding of disease evolution, and to predict the future risk of disease onset and progression.

Prediction of disease progression is particularly important in age-related macular degeneration (AMD). AMD is the commonest cause of blindness in the developed world[6]; in the US alone an estimated 148,000 adults each year progress from the early, mild form of the condition to the sight-threatening late form known as exudative AMD (exAMD)[7–9]. Once exAMD develops, sight is lost precipitously and often cannot be fully restored by current therapies, making the point of conversion from early to exAMD a critical moment in the management of this disease[14]. Exudative AMD typically affects one eye first, leaving sufferers reliant upon the unaffected fellow eye to maintain their quality of life. However, 20% of these patients develop exAMD in the fellow eye within 2 years of the first[10–13]. This deprives individuals of essential daily activities such as reading, recognising faces, and driving.

Treatment of exAMD is most effective if administered soon after conversion[14]. Regular follow up is thus the standard of care, but is not always available[15]. While studies are exploring preventative strategies[16-17], robust methods of identifying exAMD onset prior to conversion are needed to avoid the administration of costly, invasive treatment to the fellow eyes of all patients with unilateral exAMD, many of whom will never develop late disease in their fellow eye. To date there has been little evidence that clinicians are able to accurately predict a patient's imminent conversion, and despite progress in deriving prognostic indicators from fundus imaging[18], further work is needed to achieve clinically useful predictive accuracy.

To address this challenge we introduce an AI system to predict whether a fellow eye will convert to exAMD imminently, defined as within the ensuing 6 month period, using optical coherence tomography scans (OCT). To demonstrate the clinical applicability of this system, we explore its use across varying operating points (sensitivity/specificity pairs), and investigate the number and potential patient impact of false positive outputs. To better understand expert performance on the task, we compare model performance with Retinal Specialists and Optometrists in a benchmark study using a defined *silver* standard of conversion date. We further investigate automatic segmentations of clinically relevant tissue types to identify early changes and study high-risk subgroups.

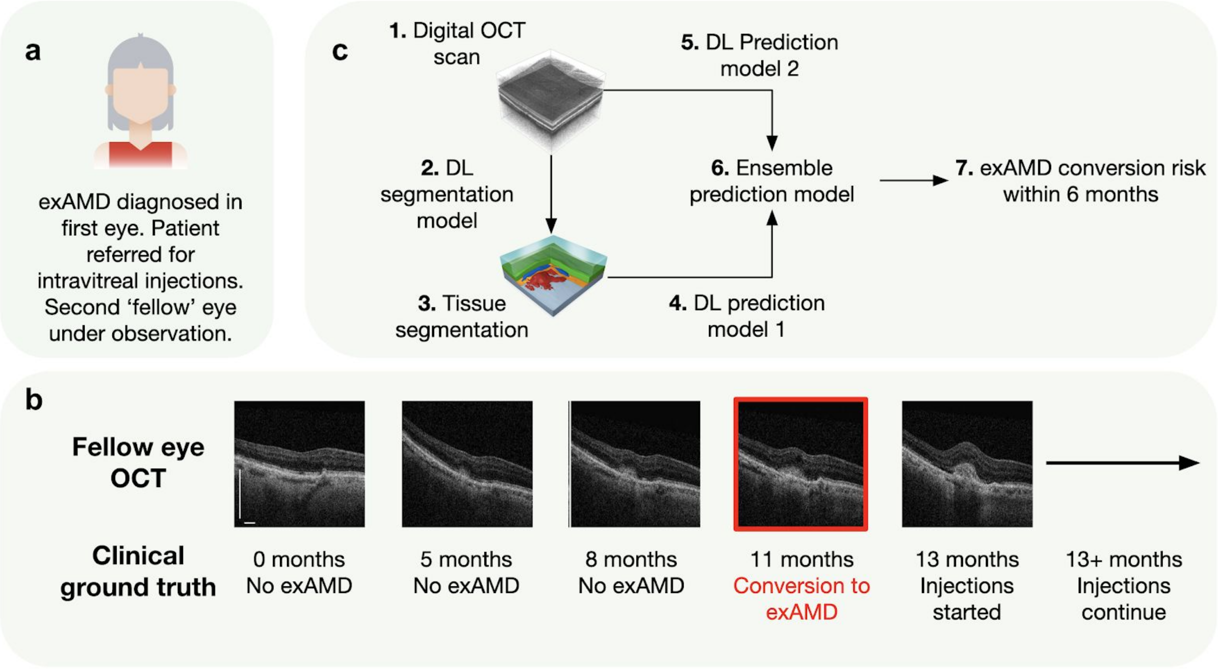

a
exAMD diagnosed in first eye. Patient referred for intravitreal injections. Second 'fellow' eye under observation.
c
1. Digital OCT scan
2. DL segmentation model
3. Tissue segmentation
4. DL prediction model 1
5. DL Prediction model 2
6. Ensemble prediction model
7. exAMD conversion risk within 6 months
b
Fellow eye OCT
Clinical ground truth
0 months No exAMD
5 months No exAMD
8 months No exAMD
11 months Conversion to exAMD
13 months Injections started
13+ months Injections continue

Figure 1 | **Clinical setup and proposed system.** a) After diagnosis of exudative AMD (exAMD) in one eye (the first eye), a patient commences intravitreal therapy in that first eye. Both the first eye and the second, 'fellow' eye are followed-up regularly with further observation. b) Selected sequential scans from the fellow eye of a patient. This eye initially showed mild, early AMD and then converted to exAMD, following which it was treated with intravitreal therapy. The timing of each follow-up visit varies depending on the treatment regimen of the first eye as well as factors related to the individual patient and the clinic. At each visit, an optical coherence tomography (OCT) scan of the first eye is performed to assess efficacy of treatment. An OCT scan of the fellow eye is also performed, as the presence of exAMD in one eye presents a high risk of fellow eye conversion. Here, the fellow eye converts to exAMD during follow up at 11 months (red box). c) Illustration of the proposed AI system. The 3D OCT volume of the fellow eye (1) is used to provide a risk prediction of whether the eye will convert within a given time-window. A deep learning (DL) segmentation model (2) outputs a 3D segmentation of anatomical and pathological tissue (3). A prediction model then takes this tissue segmentation as an input (4). A further prediction model takes the original 3D OCT volume as an input (5), and these two prediction models are ensembled (6) to assign a risk of conversion to exAMD within a clinically-actionable time window of 6 months (7).

Results

Clinical Application & Deep Learning Architecture

Predicting the future state of a progressive disease is a combination of two skills: identification of subtle signs early in the process of conversion, and modeling the future risk of exAMD. A model must be able to provide interpretable information to clinicians who may be making decisions based on its predictions. Our proposed system thus consists of two components: first predicting conversion to exAMD based on an interpretable tissue segmentation of the OCT and second making a prediction based on the raw OCT itself. The former adapts a two stage architecture[2], trained on a subset of manually segmented scans, that first segments 13 relevant tissue types, and subsequently applies a classification network adapted to predict the risk of conversion to exAMD within the next 6 months. We ensembled the two stage network with a model trained for the same task on the raw OCT alone. This was motivated by literature precedent on involvement of imaging features not yet captured by the segmentation model such as reticular pseudodrusen[19–22] and reflectivity of tissues[23] in the conversion to exAMD. This approach captures the complementary performance of the two stage network and those based on raw OCT alone (**Supplementary Tables 1a & 1b**).

We trained and tested our system using a retrospective, consecutive cohort of 2,795 patients across seven different sites (**Supplementary Table 2**) who were first diagnosed with exAMD between June 2012 and June 2017 (**Figure 1**). Routine care for these patients comprises repeated bilateral OCT at varying intervals, most commonly every 4-12 weeks whilst undergoing therapy, and every 3-12 months if therapy is ceased to monitor for disease reactivation. The dataset consisted of 62% female and 38% male patients. Ethnicities were 55% Caucasian, 10% Asian, 2% Black, and 33% other or unknown. Average age at first eye presentation was 78.8 years (see **Extended Data Figure 1**). These figures reflect the epidemiology of AMD[24].

Fellow eyes were grouped into converting and non-converting within the follow-up available, referring to whether they converted to exAMD during the study period. All patients had a follow up period for their fellow eye of ≥6 months. To account for potential differences between the date a fellow eye converted, and when therapeutic injections were started, all scans underwent expert review to provide a clinical ground truth of a conversion scan in addition to the injection scan. The mean and median differences between these two events were 64.9 days and 13 days respectively (**Extended Data Figure 1, Extended Data Figure 2**). The dataset was randomly split at the patient level into model training and validation sets

(80%) and a hold-out test set on which to evaluate final model performance (20%) (see **Supplementary Table 3** and **Online Methods**).

Future prediction of conversion to exudative AMD

We evaluated our model on a primary outcome of identification of OCT scans at risk of developing exAMD within the ensuing 6 months. We chose this time window to enable the model to predict at least two follow-up intervals ahead of time, assuming a maximal follow-up interval of 3 months.

Our system reached a per volumetric-scan area under the receiver operating characteristic curve (AUC) of 0.745 on the test set, predicting the clinical ground truth of 'conversion scan', and an AUC of 0.884 when compared to a ground truth of the actual injection date (**Figure 2**). All following results use the 'conversion scan' ground truth. This substantially outperformed baseline Gradient Boosted Machine models trained only on available demographic metadata (**Supplementary Table 4**).

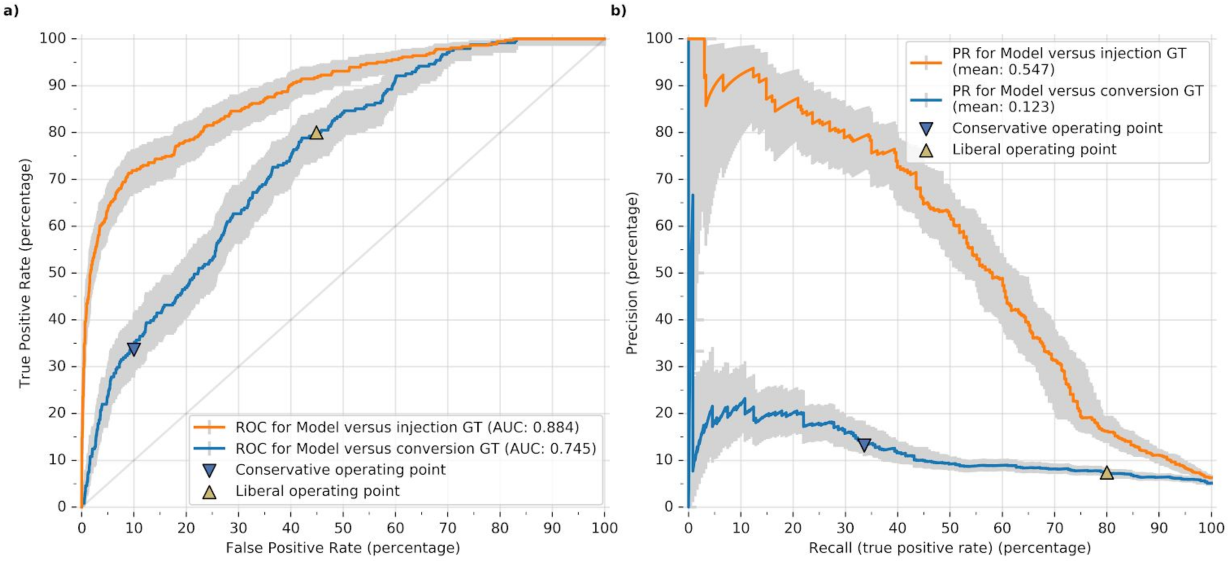


Figure 2 | Results on prediction of conversion within 6 months with conversion and first injection as ground truth. a) Receiver operating characteristic (ROC) curves showing per-volumetric scan performance on prediction of imminent conversion within 6 months on the full test set (386 unique patients; 5,581 total OCT scans) against a ground truth (GT) defined as the first injection occurring within 6 months, in blue (N(positive cases)=363, N(negative cases)=6578); and against a ground truth defined as the conversion date, in orange (N(positive cases)=241, N(negative cases)=5340). Conservative (90% specificity) and liberal (80% sensitivity) operating points shown. The numbers of true and false cases differ due to the exclusion of scans with unknown conversion dates. Gray shaded areas indicate 95% confidence intervals (see Methods). The grey diagonal line indicates chance performance. b) Precision-recall (PR) curves on the same results.

Predicting future conversion is not a routine clinical task, and there are many ways in which our proposed AI system could be used in practice. One way of exploring this is through the balance between sensitivity and specificity, where changes in management such as visit scheduling may require different operating points compared with currently unproven preventative treatments. Without risk prediction, a conservative approach would entail providing the same management to every patient with exAMD in one eye, the highest possible false positive rate. This may not be practical, and in current practice there is no provision for changes in management for those most at risk of progression. To represent this balance, we propose a conservative (90% specificity) and a liberal (80% sensitivity) operating point. At the conservative point a sensitivity of 34% is achieved at 90% specificity; at the liberal point a specificity of 55% is achieved at 80% sensitivity. This corresponds to false positives in only 9.6% of scans at the conservative operating point, and 43.4% of scans at the liberal operating point. There was minimal difference in the true or false positive rates at the conservative or liberal operating points when weighing the AUC to balance the average number of scans per patient (**Supplementary Table 5**).

This approach could be extended across varying lead-times and a range of different operating points as required by individual clinics, healthcare systems, or therapeutic drug indications. **Extended Data Figure 3** and **Supplementary Table 6** give examples of such extensions.

To better reflect the alternative ways in which this system could be applied we explore the performance for individual patients (n=386) rather than scans. If applied in practice, a single correct positive prediction is sufficient to begin a potentially beneficial course of treatment if preventative treatment commenced. In patients whose fellow eyes converted during the study period (n=103), the system produced true positives in at least one scan during the preceding 6 months in 40.8% and 77.7% of the converting eyes for the conservative and liberal operating points respectively. Conversely, a false positive alert could lead to unnecessary treatment. For fellow eyes with at least 6 months of negative follow-up (n=386), the system produced at least one false positive in 23.1% of individuals at the conservative operating point, and 61.1% at the liberal operating point. Considering only those with a longer follow-up of 24 months (n=208), this dropped to 16.8% and 55.8% for the conservative and liberal operating points respectively (**Supplementary Table 7a**).

Patients can still be managed effectively outside the 6 month window if it is expected that a patient will convert to exAMD. We investigate false positive predictions in fellow eyes where conversion did not occur within the 6 month window but later in the patient's clinical history. At the conservative and liberal operating points, 23.6% and 25.8% respectively of all fellow eyes with false positive predictions were 'early' and converted greater than 6 months after the initial model prediction. For patients with a follow-up of at least 24 months after initial model prediction, we investigate the number of false positive alerts that converted within a 6-24 month period. At the conservative and liberal operating points this was 35.2% and 32.8% respectively (**Supplementary Table 7b**).

Clinical expert benchmark for future prediction

Predicting future conversion to exAMD is not a routine task performed by clinicians. In current practice, scans are assessed for signs of having already converted. Though several of prognostic imaging features have been described[25], clinical expert performance at prediction of future fellow eye conversion has not previously been studied.

It is essential to establish a benchmark for human performance in practice to understand the performance of our proposed system. We used an enriched subset of the test set to meet statistical power, randomly choosing at least one scan in the 6 months prior to conversion for each converting fellow eye, resulting in a prevalence of 13.5% of scans that converted within 6 months (see **Online Methods**). For each case, we obtained the predictions from our system and six clinical experts: three Retinal Specialists and three Optometrists trained in medical retina. Each expert was asked to predict whether the eye would convert in the following 6 months, and provided two separate decisions at least one week apart: one (like our system) from a single OCT scan (single scan task) (**Dataset #9 in Supplementary Table 8**), and one from the OCT with available historical OCTs, fundus images and patient demographic and visual acuity data (sequential scan task) (**Dataset #10 in Supplementary Table 8**). We compared these against the clinical ground truth of time to exAMD conversion.

Despite the task not being routinely performed by clinicians, the experts performed better than chance alone. However, performance varied substantially; sensitivity ranged from 18-56%, and specificity from 61-93% for the single scan task. On average, when given additional information specificity improved at the expense of reduced sensitivity (sensitivity range 8.5-41.5%, specificity range 77.4-98.6%). Inter-rater agreement for the single scan task was slightly better among Retinal Specialists (κ=0.335) than

Optometrists (κ=0.258). For the sequential task, agreement between the Retinal Specialists (κ=0.143) was worse than the Optometrists (κ=0.305). Intra-observer agreement between the single and sequential scan tasks ranged between a κ of 0.180 and 0.523. Further details on individual expert decisions are given in **Extended Data Figure 4**, and additional expert metrics and agreement comparisons are given in **Extended Data Figure 5.**

Comparing these results for future prediction to our system, we outperform the majority of experts (**Figure 3**). The system had a higher performance than five experts (all three Retinal Specialists, two Optometrists) and matched one (an Optometrist) for the single scan task. When experts additionally had access to each patient's previous OCT scans, fundus images and additional clinical information, our model again outperformed five experts (two Retinal Specialists and all three Optometrists), while one (a Retinal Specialist) was similar to our system. The system achieves a significantly better F1 score at the equal error point compared to five out of six experts (model, 0.38, human experts, 0.23-0.33; **Supplementary Table 9**). We evaluate the conservative and liberal operating points with a McNemar test between each expert and the points (**Supplementary Table 10**). At the conservative operating point the model has significantly greater sensitivity than 3 experts, and significantly greater specificity than 2 experts. At the liberal operating point, where we trade specificity for sensitivity, the model has, as expected, a significantly better sensitivity than all experts but with a significantly worse specificity.

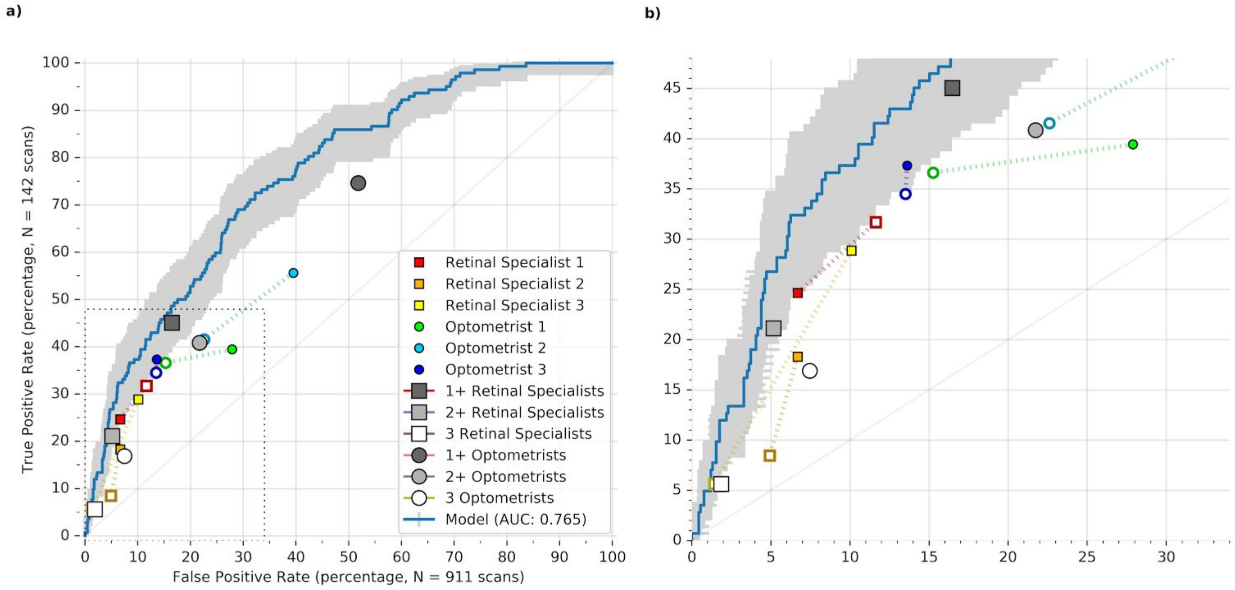


Figure 3 | **Results from the clinical benchmark study.** a) ROC curve showing the performance of the AI system on the clinical expert benchmark subset. Clinical experts are represented by filled circles for the single scan task and open circles for the sequential scan task. The individual points for single and sequential scan tasks for each expert are linked by dashed lines. The larger monochrome squares and circles show a human performance where a prediction of conversion requires at least n or more (n+) retinal specialists or optometrists in agreement the case will convert in the next 6 months. Grey shaded regions areas indicate 95% confidence intervals (see Methods). The grey diagonal line indicates chance performance. b) Close-up of the region between 0-30% false positive rate (outlined as a dotted region in (a)).

Visualising anatomical subgroups

Additional information that is interpretable to clinicians can aid effective implementation[26]. One such benefit of our system is that it automatically segments each scan. Extracting clinically relevant features provides a systematic method to visualise change over time (**Extended Data Figure 6**). **Figure 4** shows a representative example, combining the risk predictions with top down two-dimensional *enface* maps created from the automatic 3D segmentations. Further examples are provided in **Supplementary Figures 1-7**.

By enabling the visualisation of important anatomy and pathology, segmentations also provide a quantitative method to derive clinical subgroups based on segmented tissue volumes (**Table 1, Supplementary Table 11**). One clinically relevant example is provided by the Age-Related Eye Disease Study (AREDS) Simplified Severity Scale used to assess 5-year conversion risk in clinical practice from fundus photographs, based on the size of the drusenoid pigment epithelial detachment (PED)[27]. Taking advantage of the 3D nature of OCT we approximate this scale using drusen volume. The findings were consistent with literature precedent - higher conversion rates were seen in subgroups with greater drusen volume. This approximation can also serve as a baseline with which to compare the model, imitating the existing scenario where AREDS has been used for recruitment into clinical trials of prophylactic treatments for exAMD. Our model outperforms measures based on drusen or hyper-reflective foci (HRF) alone (**Figure 5**).

This approach provides insight into model performance. The system is substantially more sensitive when features known to be predictive, such as HRF [28-29] and high drusen volumes[30-31], are present (**Table 1**). This is also the case for fibrovascular pigment epithelial detachment (PED) present prior to conversion, possibly highlighting early exudative changes. We show the system's performance is consistent across key demographics such as sex and ethnicity, provide more examples of clinically important subgroups including cases selected by the appearance of the conversion scan in **Supplementary Tables 12 & 13,** and present Kaplan-Meier survival plots for individual subgroups in **Extended Data Figure 7**.

Table 1 | System performance for selected examples of patient subgroups, as identified using automatic segmentation. All subgroups were derived using automated segmentation of individual volumetric OCT scans of individual eyes, taken from the test set, that did not show exAMD on first presentation.

Patient subgroup	Number of scans	Imminency scan prevalence (%)	Imminency AUC (95% CI)	Conservative operating point (90% specificity)		Liberal operating point (80% sensitivity)	
				Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)
All patients	5581	4.3	0.745 (0.718-0.772)	33.6	90.0	79.7	55.1
No drusen (no AMD)	425	0.0	n/a	N/A	100.0	N/A	95.1
Drusen volume 0-25th percentile	971	0.5	0.915 (0.862-0.968)	0.0	97.8	60.0	88.4
Drusen volume 25-50th percentile	1395	4.3	0.681 (0.616-0.746)	15.0	94.8	68.3	61.9
Drusen volume 50-75th percentile	1395	5.9	0.584 (0.515-0.649)	24.1	87.3	72.3	36.3
Drusen volume 75-100th percentile	1395	6.7	0.759 (0.710-0.808)	55.9	78.8	94.6	29.6
Geographic atrophy present	1573	4.0	0.692 (0.633-0.751)	31.7	85.7	88.9	34.9
Geographic atrophy absent	4008	4.4	0.774 (0.742-0.806)	34.3	91.7	76.4	63.1
HRF present	3867	5.3	0.725 (0.692-0.758)	38.8	86.7	87.9	41.7
HRF absent	1714	2.0	0.779 (0.737-0.821)	2.9	97.3	31.4	84.6
Fibrovascular PED present	2326	6.6	0.675 (0.632-0.718)	48.1	77.3	90.9	22.4
Fibrovascular PED absent	3255	2.7	0.784 (0.746-0.822)	8.0	98.7	59.8	77.6

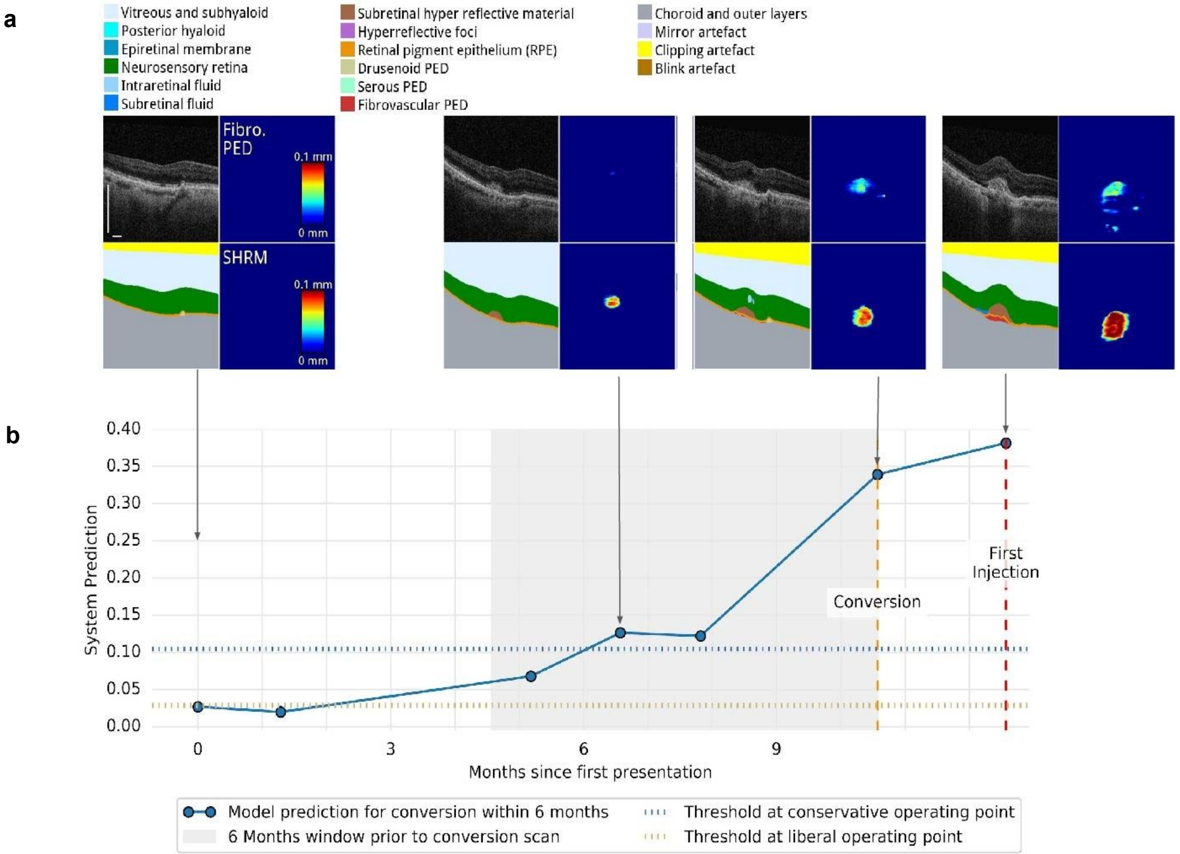


Figure 4 | Example of a correct prediction by the AI system. In this example, progression of the right 'fellow' eye of an 80-year-old male patient being treated in the left eye for exudative AMD with intravitreal injections is shown. The patient was seen at regular intervals, over which time his right eye showed a gradual progression in anatomical abnormalities, before converting to the exudative form approximately 11 months after first presentation and receiving therapeutic injections beginning at 13 months. (a) For each set of images, shown are B-scan slices of the OCT imaging (top left), the segmentation produced by a DL segmentation model2 (bottom left), and en face thicknesses of two clinically important retinal tissues produced by the segmentation model, where blue=0mm and red=0.1mm (fibrovascular pigment epithelial detachment (fibro. PED, top right) and subretinal hyper-reflective material (SHRM, bottom right)). Each set of four images is from a clinical visit (selected visits indicated with arrows) during the 12-month period shown. In the months leading to conversion, we observed an increasing presence of SHRM and fibrovascular PED. At 10.5 months, the volume of SHRM and fibrovascular PED increased further and intraretinal fluid was observed (en face map not shown), signaling conversion to exAMD. Treatment commenced 2 months later; at this point further anatomical changes had occurred, including an additional OCT finding of subretinal fluid (en face map not shown). b) Prediction of the AI system for conversion to exAMD within 6 months. At the liberal operating point (yellow dotted line) it correctly predicted conversion within 6 months for all three scans within the actual 6 month window prior to conversion (grey box).

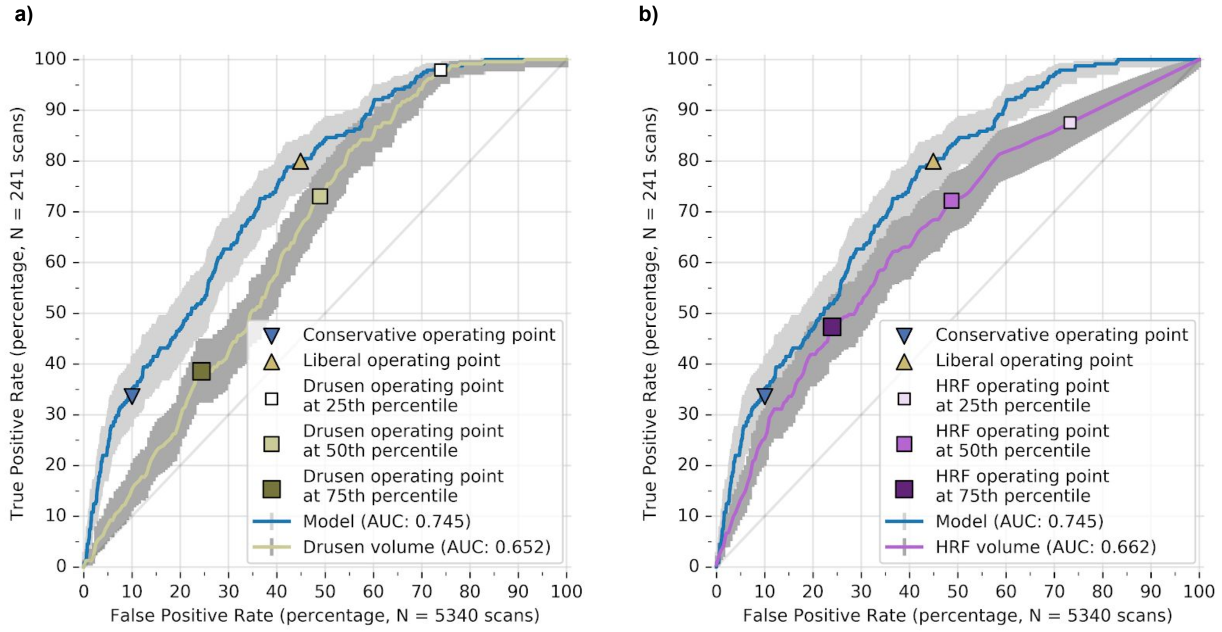


Figure 5 | AUC curve comparisons of the system against benchmark predictions based on the volume of drusen and hyper-reflective foci (HRF). (a,b) Estimation of the performance of the Age-Related Eye Disease Study (AREDS) Simplified Severity Scale in classifying patients as ‘high risk’. Shown are ROC curves using the volume of drusen (a) and of hyper-reflective foci (HRF) (b). The curves were built by varying the volume threshold at which a prediction of 6-month conversion would be triggered and evaluating this against the ground truth. Operating points are shown at the 25th, 50th and 75th percentiles of volume.

Discussion

We demonstrate an AI system to predict conversion to exAMD in fellow eyes of patients with exAMD in their first eye. We propose two clinically applicable operating points and consider the system's potential impact to clinical care through analysis of false positive alerts, and demonstrate the value of automatic segmentation in identifying early signs of progression and studying high-risk subgroups. We establish a clinical expert benchmark for comparison as this task is not currently performed in clinical practice, showing humans are able to perform the task albeit with high variability.

Our system has several potential implications for patient care. Future prediction of conversion is important in guiding preventative measures in AMD. These are already being explored, with examples of prophylactic intravitreal therapy[16-17] and gene therapies[32] in clinical trials, and longer-acting intravitreal therapy[33,34] and port-delivery systems for long-term continuous therapy[35] under investigation. Means of identifying those most at risk are required if these therapies are to be efficiently targeted in increasingly burdened healthcare systems and are to be acceptable to patients. Our proposed system outperforms volume based risk predictions similar to those currently used for trial cohort selection, and may enable targeting of preventative treatments and identification of high-risk patients for inclusion in similar upcoming trials. The operating points are configurable and will vary depending on the use, healthcare system and therapy of choice.

A further implication for care is in influencing patient follow-up and improving time to treatment. Early diagnosis is paramount as delays in intervention can result in a loss of vision[36]. However, the mean difference between injection date and conversion date in our dataset was 64.9 days (median, 13 days). One explanation for this gap is that subtle early signs are not always being treated because they are missed or do not fall within set treatment criteria, or because the patients were asymptomatic. This only partially explains the difference: there is still a substantial delay, a mean of 34 days, even in those within treatment criteria. The model may be particularly beneficial in these cases. In addition to predicting conversion, the segmentations produced provide information aimed at earlier detection of exAMD. The improvement in system performance when trained and evaluated using the scan from the date of injection as a ground truth further indicates the potential to identify conversion changes earlier. Moreover, our system does not require sequential information. While including a patient's previous scans and demographics led to mixed results for the experts, it is plausible that AI can extract additional useful information[5]. However, predicting only on single scans supports settings where patient follow up varies depending on perceived risk. This is particularly relevant in centers that cannot offer regular follow up, especially with increasing availability of OCT through community eye care centers, and for future work to investigate the applicability of our system in patients that have yet to develop exAMD in either eye.

The segmentation portion of the model enables automated detection and analysis of important tissues[2]. One use of this is to study groups of scans based on known prognostic indicators from the OCT, as well as other important phenotypes such as the pathological tissues present on the conversion scan. Not only do the *enface* maps provide summary information to clinicians treating a patient, but they may open up new ways to study AMD subgroups, and indeed other conditions, which may differ in their conversion risk or response to treatment in important ways.

Further imaging in patients where the model produces false positive predictions may demonstrate particular subgroups of interest. Newer imaging modalities such as OCT angiography (OCTA) are becoming more widely used to safely acquire high-resolution images of the choroidal vasculature to identify exAMD. Recent studies employing OCTA have distinguished a form of exAMD coined as *subclinical* or *non-exudative* neovascular AMD that does not result in the appearance of macular fluid visible on conventional OCT or modalities such as fundus fluorescein angiography[37-39]. For our study, the clinical ground truth of conversion was labeled where exudation was visible. Eyes with suspected fibrovascular PED without the presence of fluid were labeled as non-exudative, but may represent examples of subclinical exAMD. It is possible that our model has identified examples that would warrant further imaging using OCTA (**Supplementary Figure 6**). Such patient groups would offer an explanation for both early findings of fibrovascular material prior to conversion (6.6% of immenancy scans), and the clustering of false positive alerts in a small number of cases. Further work is needed to validate this hypothesis by evaluating model predictions with OCTA.

Our work builds upon a body of literature investigating the development of AMD[40], and promising early work to develop predictive models for exAMD based on fundus photographs[41] and OCT scans[42–45]. We improve on generalisability and applicability in several ways. Our datasets are representative of the patient population at a large specialist eye-care centre, both in the cadence of patient visits and the inclusion of challenging cases, such as eyes with geographic atrophy (25.2% of eyes with geographic atrophy converted to exAMD in our dataset). Crucially, our clinical ground truth reflects the date of conversion rather than using injection as a proxy measure. There is often a delay between conversion and treatment; when first injection date is used as the conversion label for training model performance improves substantially. In addition, when the ground truth is based on the injection, exAMD masquerades are mislabelled as they may still receive injections.

Clinicians do not routinely make predictions about future conversion. Our results indicate that clinical experts are able to perform the task, but with large variability. Though specificity improved for all experts when given the full clinical scenario with all historical images and additional patient information, the sensitivity reduced. While an exploration of variability is beyond this work's scope, our results open up the possibility of exploring human performance on this task as has been investigated in conditions such as diabetic retinopathy[46]. The low inter-rater agreement between individual experts may reflect that predicting conversion is not routine. Despite some literature evidence of prognostic features, no formal prognostic criteria exists. Standardised training can reduce variability between individuals, but without established criteria such training is impossible. It is possible that models may reduce this variability; future work can investigate this through human-computer interaction studies.

There are several limitations of our work. While our system was trained and evaluated on a diverse and clinically representative demographic from Moorfields Eye Hospital, they are not fully representative of a global population. AMD is multifactorial, with genetics, race, sex, and lifestyle factors such as smoking and diet known to contribute to disease risk[47]. Its incidence varies globally, being lower in Asia and Africa compared to Europe and North America[48]. Additional representative datasets would be required to confirm performance on a general population. In addition we only test our model on one OCT scanner type. Different models may vary in appearance; future work should investigate generalisability across OCT manufacturers. We powered the study based on and report performance across individual scans. While we explore subgroups of the full dataset for per-patient and segmentation analyses, the statistical power is limited and future work should include larger datasets. We investigate a 6-month time window prior to a clinical ground truth of conversion date. This clinical ground truth is defined based on an OCT scan demonstrating exudative conversion. It is unlikely the date that patient conversion corresponds exactly to when the scan was taken, but rather to a point in time between the current and previous scans. This difference may account for some of the false positives that occur in patients that do still convert outside the 6-month time window. There are differences in performance by training a model on raw OCT scans compared with training on the OCT segmentation. Though small, the differences suggest there are important imaging features for this task that are not captured by the segmentation model; further work could extend the segmentation model with the addition of a wider range of different tissue classes to improve performance[24] and investigate models using segmentation alone. Cases of undiagnosed PCV that may masquerade as positive examples of exAMD were removed by manual OCT grading. Indocyanine green angiography can be used to confirm this diagnosis, but was unavailable routinely in the Moorfields dataset. A final limitation is that there may be important differences in treatment regimes and other patient factors that correlate with the number of scans a patient has. Future work should investigate potential bias across larger datasets.

Our model was trained and evaluated on a dataset of fellow eyes of patients with exAMD in one eye. This is a population at high-risk of developing exAMD in their second eye, and associated loss of vision substantially impacts quality of life in a patient who has already lost vision in their first eye. Future work can build on these results through prospective implementation and validation studies, and by investigating model performance in patients without any AMD, or with dry AMD in one or both eyes.

In summary, we introduce a clinically applicable AI system that produces a prediction of fellow eye conversion to exAMD based on OCT scans from a clinically relevant population, and provides additional information to clinicians through automatic segmentation. The system opens up new possibilities for research and treatment for the leading cause of blindness in the developed world.

References

1. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
2. De Fauw, J. *et al.* Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **24**, 1342–1350 (2018).
3. Ardila, D. *et al.* End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* **25**, 954–961 (2019).
4. Poplin, R. *et al.* Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng* **2**, 158–164 (2018).
5. Tomašev, N. *et al.* A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* (2019).
6. Wong, W. L. *et al.* Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. *Lancet Glob Health* **2**, e106–16 (2014).
7. Owen, C. G. *et al.* The estimated prevalence and incidence of late stage age related macular degeneration in the UK. *Br. J. Ophthalmol.* **96**, 752–756 (2012).
8. Rein, D. B. *et al.* Forecasting Age-Related Macular Degeneration Through the Year 2050: The Potential Impact of New Treatments. *Arch. Ophthalmol.* **127**, 533–540 (2009).
9. Rudnicka, A. R. *et al.* Incidence of Late-Stage Age-Related Macular Degeneration in American Whites: Systematic Review and Meta-analysis. *American Journal of Ophthalmology* **160**, 85–93.e3 (2015).
10. Bek, T. & Klug, S. E. Incidence and risk factors for neovascular age-related macular degeneration in the fellow eye. *Graefes Arch. Clin. Exp. Ophthalmol.* **256**, 2061–2068 (2018).
11. Zarranz-Ventura, J. *et al.* The neovascular age-related macular degeneration database: report 2: incidence, management, and visual outcomes of second treated eyes. *Ophthalmology* **121**, 1966–1975 (2014).
12. Fasler, K. *et al.* The Moorfields AMD Database Report 2 - Fellow Eye Involvement with Neovascular Age-related Macular Degeneration. doi:10.1101/615252
13. Maguire, M. G. *et al.* Incidence of choroidal neovascularization in the fellow eye in the comparison of age-related macular degeneration treatments trials. *Ophthalmology* **120**, 2035–2041 (2013).
14. Lim, J. H. *et al.* Delay to treatment and visual outcomes in patients treated with anti-vascular endothelial growth

factor for age-related macular degeneration. *Am. J. Ophthalmol.* **153**, 678–86, 686.e1–2 (2012).

15. Amoaku, W. *et al.* Action on AMD. Optimising patient management: act now to ensure current and continual delivery of best possible patient care. *Eye* **26 Suppl 1**, S2–21 (2012).
16. IAI Versus Sham as Prophylaxis Against Conversion to Neovascular AMD (PRO-CON): ClinicalTrials.gov Identifier: NCT02462889. *ClinicalTrials.gov* Available at: https://clinicaltrials.gov/ct2/show/NCT02462889.
17. Prophylactic Ranibizumab for Exudative Age-related Macular Degeneration (PREVENT). ClinicalTrials.gov Identifier: NCT02140151. *ClinicalTrials.gov* Available at: https://clinicaltrials.gov/ct2/show/NCT02140151.
18. Chew, E. Y., Lindblad, A. S. & Clemons, T. Summary Results and Recommendations From the Age-Related Eye Disease Study. *Arch. Ophthal.* **127**, 1678 (2009).
19. Cohen, S. Y. *et al.* Prevalence of reticular pseudodrusen in age-related macular degeneration with newly diagnosed choroidal neovascularisation. *Br. J. Ophthalmol.* **91**, 354–359 (2007).
20. Zweifel, S. A., Imamura, Y., Spaide, T. C., Fujiwara, T. & Spaide, R. F. Prevalence and Significance of Subretinal Drusenoid Deposits (Reticular Pseudodrusen) in Age-Related Macular Degeneration. *Ophthalmology* **117**, 1775–1781 (2010).
21. Zhou, Q. *et al.* Pseudodrusen and Incidence of Late Age-Related Macular Degeneration in Fellow Eyes in the Comparison of Age-Related Macular Degeneration Treatments Trials. *Ophthalmology* **123**, 1530–1540 (2016).
22. Lee, J., Choi, S., Lee, C.S., Kim, M., Kim, S.S., Koh, H.J., Lee, S.C. and Byeon, S.H.. Neovascularization in fellow eye of unilateral neovascular age-related macular degeneration according to different drusen types. *Am. J. Ophthalmol.* **208**, 103-110 (2019).
23. Veerappan, M. *et al.* Optical Coherence Tomography Reflective Drusen Substructures Predict Progression to Geographic Atrophy in Age-related Macular Degeneration. *Ophthalmology* **123**, 2554–2570 (2016).
24. VanderBeek, B.L., Zacks, D.N., Talwar, N., Nan, B., Musch, D.C. and Stein, J.D. Racial differences in age-related macular degeneration rates in the United States: a longitudinal analysis of a managed care network. American journal of ophthalmology, 152(2), 273-282 (2011).
25. Age-Related Eye Disease Study Research Group. A Simplified Severity Scale for Age-Related Macular Degeneration: AREDS Report No. 18. *Arch. Ophthal.* **123**, 1570 (2005).
26. Sana Tonekaboni, Shalmali Joshi, Melissa D. McCradden, Anna Goldenberg. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. *arXiv* (2019). Available at: https://arxiv.org/pdf/1905.05134.pdf. (Accessed: 2nd August 2019)

27. Age-Related Eye Disease Study Research Group. The Age-Related Eye Disease Study system for classifying age-related macular degeneration from stereoscopic color fundus photographs: the Age-Related Eye Disease Study Report Number 6. *Am. J. Ophthalmol.* **132**, 668–681 (2001).

28. Fragiotta, S., Rossi, T., Cutini, A., Grenga, P. L. & Vingolo, E. M. PREDICTIVE FACTORS FOR DEVELOPMENT OF NEOVASCULAR AGE-RELATED MACULAR DEGENERATION: A Spectral-Domain Optical Coherence Tomography Study. *Retina* **38**, 245–252 (2018).

29. Schmidt-Erfurth, U. *et al.* Prediction of Individual Disease Conversion in Early AMD Using Artificial Intelligence. *Invest. Ophthalmol. Vis. Sci.* **59**, 3199–3208 (2018).

30. Abdelfattah, N. S. *et al.* Drusen Volume as a Predictor of Disease Progression in Patients With Late Age-Related Macular Degeneration in the Fellow Eye. *Invest. Ophthalmol. Vis. Sci.* **57**, 1839–1846 (2016).

31. Folgar, F. A. *et al.* Drusen Volume and Retinal Pigment Epithelium Abnormal Thinning Volume Predict 2-Year Progression of Age-Related Macular Degeneration. *Ophthalmology* **123**, 39–50.e1 (2016).

32. NIHR Oxford Biomedical Research Centre. World’s first gene therapy operation for common cause of sight loss carried out. (2019).

33. Dugel, P. U. *et al.* HAWK and HARRIER: Phase 3, Multicenter, Randomized, Double-Masked Trials of Brolucizumab for Neovascular Age-Related Macular Degeneration. *Ophthalmology* (2019). doi:10.1016/j.ophtha.2019.04.017

34. Sahni, J. *et al.* Simultaneous Inhibition of Angiopoietin-2 and Vascular Endothelial Growth Factor-A with Faricimab in Diabetic Macular Edema: BOULEVARD Phase 2 Randomized Trial. *Ophthalmology* **126**, 1155–1170 (2019).

35. Campochiaro, P. A. *et al.* The Port Delivery System with Ranibizumab for Neovascular Age-Related Macular Degeneration: Results from the Randomized Phase 2 Ladder Clinical Trial. *Ophthalmology* (2019). doi:10.1016/j.ophtha.2019.03.036

36. Muether, P.S., Hermann, M.M., Koch, K. and Fauser, S.. Delay between medical indication to anti-VEGF treatment in age-related macular degeneration can result in a loss of visual acuity. Graefe's Archive for Clinical and Experimental Ophthalmology, 249(5), 633-637 (2011).

37. Roisman, L. *et al.* Optical Coherence Tomography Angiography of Asymptomatic Neovascularization in Intermediate Age-Related Macular Degeneration. *Ophthalmology* **123**, 1309–1319 (2016).

38. de Oliveira Dias, J. R. *et al.* Natural History of Subclinical Neovascularization in Nonexudative Age-Related

Macular Degeneration Using Swept-Source OCT Angiography. *Ophthalmology* **125**, 255–266 (2018).

39. Carnevali, A., Sacconi, R., Querques, L., Marchese, A., Capuano, V., Rabiolo, A., Corbelli, E., Panozzo, G., Miere, A., Souied, E. and Bandello, F., 2018. Natural History of Treatment-Naïve Quiescent Choroidal Neovascularization in Age-Related Macular Degeneration Using OCT Angiography. Ophthalmology Retina, 2(9), pp.922-930.

40. Jager, R. D., Mieler, W. F. & Miller, J. W. Age-related macular degeneration. *N. Engl. J. Med.* **358**, 2606–2617 (2008).

41. Babenko, B. *et al.* Predicting Progression of Age-related Macular Degeneration from Fundus Images using Deep Learning. (2019).

42. Schmidt-Erfurth, U. *et al.* Prediction of Individual Disease Conversion in Early AMD Using Artificial Intelligence. *Invest. Ophthalmol. Vis. Sci.* **59**, 3199–3208 (2018).

43. Bogunovic, H. *et al.* Machine Learning of the Progression of Intermediate Age-Related Macular Degeneration Based on OCT Imaging. *Invest. Ophthalmol. Vis. Sci.* **58**, BIO141–BIO150 (2017).

44. Russakoff, D. B., Lamin, A., Oakley, J. D., Dubis, A. M. & Sivaprasad, S. Deep Learning for Prediction of AMD Progression: A Pilot Study. *Invest. Ophthalmol. Vis. Sci.* **60**, 712–722 (2019).

45. Banerjee, I. *et al.* A Deep-learning Approach for Prognosis of Age-Related Macular Degeneration Disease using SD-OCT Imaging Biomarkers. *arXiv* (2019).

46. Krause, J., Gulshan, V., Rahimy, E., Karth, P., Widner, K., Corrado, G.S., Peng, L. and Webster, D.R. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. Ophthalmology, 125(8), 1264-1272 (2018)

47. Vander, J. F. Risk Factors for the Incidence of Advanced Age-Related Macular Degeneration in the Age-Related Eye Disease Study (AREDS): AREDS Report No. 19. *Yearbook of Ophthalmology* **2006**, 119–121 (2006).

48. Wong, W. L. *et al.* Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. *Lancet Glob Health* **2**, e106–16 (2014).

Acknowledgements

We thank the patients under the care of Moorfields Eye Hospital. We would also like to thank B. Romera-Paredes, O. Ronneberger, N. Tomasev, S. Blackwell, J. Schrouff, M. Chesus, C. Cooper, V. Cornelius, A. Khawaja, R. Ahamed, T. Corkett, R. Ogbe, Y. Ibitoye, M. Bawn, J. Besley, C. Meaden, C. Chorley, S. Rowley, C. Semturs, A. Varadarajan, B. Babenko, I. Traynis, Y. Liu, L. Peng, N. Hammel, K. Blumer, K. Kavukcuoglu, S. Bouton, G. Corrado, E. Manna, A.C. Bird, W. Tucker, Y. Obadeyi, Z. Jessa, D. Sim, M. Natkunarajah and A. Jindal.

P.A.K. is supported by an NIHR Clinician Scientist Award (NIHR-CS-2014-14-023). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. R.C. receives studentship support from the College of Optometrists, United Kingdom.

Author contributions

R.C., P.T.K., D.K., D.H., M.S., T.B., J.R.L., P.A.K., J.D.F. initiated the project.

J.Y., R.C., T.S., H.A., M.L., J.H., K.F., G.M., J.R.L., P.A.K., J.D.F. created the dataset.

J.Y., T.S., J.W., H.A., M.W., J.D., J.D.F. contributed to software engineering.

J.Y., R.C., T.S., J.W., C.K., T.B., J.R.L., P.A.K., J.D.F. analysed the results.

J.Y., R.C., T.S., J.W., A.O., C.K., H.A., T.B., J.R.L., J.D.F. contributed to the overall experimental design

J.Y., J.W., J.D.F. designed the model architectures.

R.C., C.K., C.H., G.R., A.C.B., D.K., J.R.L., P.A.K., contributed clinical expertise.

J.Y., R.C., T.S., J.R.L., P.A.K., J.D.F. contributed to subgroup analysis experiments.

J.Y., R.C., C.K., T.S., J.D.F. contributed to statistical analysis.

J.Y., R.C., T.S., A.O., C.K., J.R.L., P.A.K., J.D.F. contributed to the human evaluation.

J.Y., R.C., T.S., J.W., C.K., M.L., J.H., M.W., J.D., J.D.F. contributed to segmentation model improvements.

J.Y., R.C., T.S., J.W., M.W., J.D., J.D.F. contributed to experiments using segmentation data.

J.Y., R.C., J.R.L., P.A.K., J.D.F. contributed to literature reviews.

J.Y., R.C., T.S., J.W., J.R.L., P.A.K., J.D.F. contributed to false positive analysis.

A.O., H.A., C.M., T.B., J.R.L., P.A.K., J.D.F. managed the project.

J.Y., R.C., G.R., J.R.L., P.A.K, J.D.F. wrote the paper.

Competing financial interests

P.A.K. and G.R. are paid contractors of DeepMind. P.A.K. has received speaker fees from Heidelberg Engineering, Topcon, Haag-Streit, Allergan, Novartis and Bayer. P.A.K. has served on advisory boards for Novartis and Bayer, and is an external consultant for DeepMind and Optos. M.L. received travel grants and a speaker fee from Bayer.

Online Methods

Ethical and IG approvals

This work, and the collection of retrospective data on implied consent, received national Research Ethics Committee (REC) approval from the Cambridge East REC and Health Research Authority approval (reference 16/EE/0253); it complies with all relevant ethical regulations. De-identification was performed in line with guidance provided by the Information Commissioner's Office's Anonymisation: managing data protection risk code of practice[49], and validated by the Moorfields Eye Hospital Information Technology and Information Governance departments respectively. Only de-identified retrospective data was used for research, without the active involvement of patients.

Further details on the methods are described in a published protocol describing the DeepMind collaboration with Moorfields Eye Hospital[50].

Datasets and Clinical Taxonomy

Dataset description

Data were collected using the Moorfields Eye Hospital electronic health record (EHR) system, querying all patients receiving intravitreal injection therapy with a diagnosis of age-related macular degeneration at 7 different Moorfields sites across London, United Kingdom. The clinical data used for the training and evaluation were collected by Moorfields Eye Hospital and transferred to DeepMind in a de-identified format. Retrospective data were aggregated from Moorfields Eye Hospital including its satellite sites where data had been archived to a central network. The data included adult patients aged over 50 years, with patients aged over 100 rounded to age 100 to retain anonymisation. Data were collected for all patients that started treatment in one eye between June 2012 and June 2017. Data for each patient were collected until June 2018, and included OCT images (acquired using Topcon 3D OCT-2000, Topcon, Japan) at every visit for both eyes where available; clinical information containing visual acuity and whether an intravitreal injection was delivered, including drug administered; and additional patient information including age in years at each scan, sex, and ethnicity. After initial exclusions this dataset consisted of 130,327 scans from 3,111 patients, and included a total of 6,149 eyes, and 2,526 fellow eyes (**Extended Data Figure 8**). **Extended Data Figure 7** shows for all fellow eyes a survival curve of conversion to exAMD from baseline.

The data was randomly divided at a patient-level into training (60%), validation (20%) and test (20%) sets. Cross-validation was performed after merging the training and validation sets (80%). **Extended Data Figure 1** contains an overview of patient demographics and the data as well as prevalence of fellow eye involvement. Further description of the dataset and labelling is provided in **Extended Data Figure 8 and Supplementary Tables 4 & 9.**

Clinical Taxonomy

All patients included in the dataset were diagnosed with exAMD in at least one eye - considered to be the first eye. If both eyes presented with exAMD, both eyes were considered as first eyes and excluded from the test set. Where there was only one first eye, the other eye without exAMD was known as the ‘fellow’ eye. Fellow eyes that developed exAMD during our study period were labelled as converting fellow eyes (see following section for the diagnosis procedure), whereas those that did not develop the condition were labelled as non-converting fellow eyes. As all patients were undergoing treatment in at least one eye, OCTs of both eyes were acquired at regular intervals - generally between 4-12 weeks depending on the drug being administered and the treatment response. The drug and treatment regime followed was either ranibizumab or aflibercept, on either pro re nata or 'treat and extend' schemes.

Clinical labelling

Once data were transferred, a labelling procedure was followed to (i) exclude eyes that were incorrectly coded as exAMD in the EHR and presented with other vascular conditions such as exudative choroidal neovascularisation (CNV) secondary to myopia, idiopathic polypoidal choroidal vasculopathy, or macular oedema, and (ii) to label the conversion scan of fellow eyes if exAMD has developed. The latter was required as a delay between conversion to exAMD and treatment was frequently observed (**Extended Data Figure 2b)**, often related to further investigations being undertaken, or if the eye was not within eligibility criteria to receive injections.

A consensus definition of conversion from OCT images has not previously been described in the literature. Hence, the clinical ground truth of conversion was defined as requiring both (1) the presence of subretinal or intraretinal fluid with (2) a suspicious pigment epithelial detachment (PED), haemorrhage or subretinal hyperreflective material (SHRM). A further definition of the presence of retinal angiomatous proliferation (RAP) with surrounding intraretinal fluid was also taken. The procedure was followed to label the first scan showing signs of conversion of fellow eyes that received treatment, by two retinal specialists and one optometrist trained in OCT interpretation. Disagreements were found in 16% of scans, and arbitrated by a senior expert, independent of the original three labelers but with knowledge of their labels, whose decision superseded. Two of the three experienced graders confirmed that untreated fellow eyes did not develop exAMD, and that first eyes were correctly diagnosed. These eyes were arbitrated by the senior expert where an untreated eye was thought to have converted to exAMD, where the diagnosis was equivocal (see **Supplementary Figure 8**), or if the eye was being considered for exclusion. As other imaging modalities often used in clinical practice such as fundus fluorescein angiography were unavailable (as it is not routinely performed in fellow eyes), and the lack of a consensus definition of conversion, we describe the OCT-derived conversion label as a *silver* standard. From this process, we found 85 fellow eyes converted and did not yet receive treatment. A number of eyes were excluded from the dataset, including 252 eyes diagnosed with other retinal disorders, and 103 eyes that started treatment at a visit without any signs of exAMD on the OCT scan. These eyes often presented with features commonly mistaken for exAMD such as vitelliform lesions, non-neovascular drusenoid PEDs with overlying fluid[51-52], and non-exudative detachments of the neurosensory retina[53]. Some examples of excluded eyes are shown in **Supplementary Figure 8**. Two patients had a fellow eye excluded due to disorders other than retinal conditions such as anterior eye conditions that obscured posterior segment imaging. These eyes were labelled prior to transfer. In addition, images were manually excluded if they were poor quality (where the major retinal interfaces were not visible), or contained significant blink or foldover artefact that obscured the relevant features described above, and would prevent a clinical decision being made.

The initial dataset for patients with confirmed nAMD in one eye consisted of 3,111 patients, 6,149 eyes, 2,526 fellow eyes, and 130,327 scans. The final dataset after exclusions were applied consisted of 2,795 patients, 4,729 eyes (including both first and fellow eyes in the the training set), 2,261 fellow eyes (777 converting and 1,484 non-converting), 96,111 OCT scans (65,633 scans after conversion to exAMD as defined by the silver standard), and 30,478 scans before conversion or without exAMD) (**Extended Data Figure 8)**. **Extended Data Figure 2** shows a histogram of the number of scans per unique eye in test and training/validation sets. A subset of patients only had first eye scans in the dataset. While all first eyes were excluded from the test set, first eyes with prior scans to conversion were included in during training to increase prevalence.

Benchmarking the expert performance

For this evaluation study, we recruited 3 Consultant Ophthalmologists with subspeciality training in Medical Retina and extensive clinical experience. These are referred to as Retinal Specialist 1, 2 and 3, with 14, 13, and 12 years of experience, respectively. Three hospital optometrists with specialist training in OCT interpretation and retinal diseases were also recruited, referred to as Optometrist 1, 2, and 3, with 14, 15 and 10 years of experience, respectively. All participants were independent and not involved in grading scans.

A subset of the test set was used for the evaluation. A stratified sample of OCT scans from the test set was selected to achieve 90% statistical power and to balance how often each eye was represented in the benchmark. For each converting fellow eye, one scan was first sampled in the 6 month period prior to conversion, and where data was available a second was sampled in the period >6 months prior to conversion. We then randomly sampled one further scan in the 6 months prior to conversion from half of the converting fellow eyes, with available scans chosen at random and independently sampled one more scan in the non-imminent period from half of the converting fellow eyes also chosen at random. Ten scans were excluded because the quality was insufficient for diagnosis. For each non-converting fellow eye, up to 3 scans were sampled conditionally on the scans having at least 6 months follow up; nine scans were excluded due to poor quality. This led to a total of a total of 1053 scans (336 eyes with 3 scans, 29 eyes with 2 scans, and 26 eyes with 1 scan), of which 13.5% converted within 6 months. Each sampling step was performed independently of the others contingent on the constraints we described.

Experts were informed that the OCT scans in the study were of untreated fellow eyes of patients with exAMD in their first eye. The primary question asked the experts to predict whether the eye will convert within the next 6 months. The experts were also given the option to select that the eye had already developed exAMD - this selection was assumed to be interchangeable for predicting that the eye will convert within 6 months. To capture the ambiguity of clinical practice, a secondary question was presented, asking the experts if the eye will convert in 6-12 months, or if the eye will *not* convert in the next 12 months.

To assess the performance in a realistic clinical environment, all scans were presented in a random order without any time constraints. The same random order was maintained for all 6 experts. The task comprised two reviews, with at least a week between them. On the initial ‘single scan’ review, only the OCT scan was presented at each trial (**Dataset #9 in Supplementary Table 8**). On second review, participants were presented with all the information available at the time of the OCT scan including all historical scans of both eyes, fundus photographs, age, sex, ethnicity, and where available information on visual acuity and treatment for both eyes (**Dataset #10 in Supplementary Table 8**). For the second review, trials were presented in a random but chronological order to avoid revealing future scans ahead of a trial that required a prediction on an earlier scan. The model only received the OCT scan.

Network Architectures and Training Protocol

Segmentation network

Previous work developed an accurate OCT segmentation network that categorises each voxel into one of 12 tissue classes and 3 different types of artefacts[2]. The network architecture was built using a three-dimensional U-Net[54]. The deep learning networks were implemented in TensorFlow[55] and Sonnet[56].

To prevent data contamination, we retrained the segmentation network from random weight initialization on the original ground truth segmentation maps while removing patients that were in the current test set. Training was performed across 300,000 training iterations with a batch size of 16 spread evenly across 16 NVIDIA Tesla V100 graphics processing units (GPU) using the TF-Replicator distributed training system[57]. All other model details, data augmentation, and training hyperparameters were kept the same as those used in De Fauw et al. (2018)[2].

In addition, a further sample of scans with dry AMD were manually segmented to increase the variety in AMD phenotypes seen by the network for training (**Dataset #4, Supplementary Table 8**). Furthermore, a new tissue class, termed 'hyperreflective foci' was added (described below). The segmentation network was trained to incorporate both of these additions. After training, the segmentation model generated predictions for every scan in the dataset including the validation and test set, providing tissue maps and volumes for 13 different tissues and 3 different types of artefact. Details of the tissue classes can be found in De Fauw et al. (2018)[2]. The segmentation maps were subsequently input into the classification or Clinical Referral model and used for Clinical Analysis (see below).

Hyperreflective foci segmentation class

Hyperreflective foci (HRF) are well-circumscribed dot or oval-shaped lesions that are present within the intraretinal layers. They can be visualised on OCT as small lesions with equal or greater hyperreflectivity than the RPE. The aetiology of these lesions vary by disease - in macular oedema, HRF often represent lipid exudates, whereas in age-related macular degeneration HRF are hypothesised to represent migrating RPE cells[58]. The presence of HRF has been associated with progression to late-stages of AMD - both geographic atrophy[59,29] and exAMD [28,29]. HRF have been shown to correlate with pigmentary changes visible on colour fundus photography[60], a feature identified in epidemiological [61,62] and clinical studies [63] as a key risk factor for AMD conversion.

As HRF is therefore likely to be a prognostic biomarker, this feature was added as a new tissue to the segmentation model. HRF in all previously manually segmented images were identified and segmented (**Dataset #3, Supplementary Table 8**). The segmentation network was subsequently retrained to predict this new tissue (**Extended Data Figure 9**).

Clinical referral and diagnosis network

The segmentation maps were used to retrain the referral and diagnosis classification model from De Fauw et al. 2018 that outputs four referral decisions and ten additional diagnoses[2]. Though the clinical referral and diagnosis task is not the focus of this study, they can be used as an auxiliary task to improve performance in the main task of exAMD prediction as discussed below. We retrained the same classification model on the current dataset where clinical diagnosis and referral labels were available - excluding any patient in our current test set. The performance of our clinical referral and diagnosis model closely matched the performance reported in De Fauw et al. 2018 (overall accuracy 94.5%). This motivated the generation of reliable distillation[64] labels by running the trained model over each scan in the dataset. These distillation labels were used as a ground truth for auxiliary tasks during training of the exAMD prediction model. We found this improved performance on the main task of future prediction.

exAMD prediction network

The prediction model learns to map an input scan in the form of a grey-scale raw OCT scan or one-hot encoded segmentation map to predictions conversion with varying lead times. Raw OCT inputs were normalized and downsampled using linear interpolation in the x and y axis while nearest-neighbor interpolation in the z-axis to prevent smoothing of subtle intensity changes across slices. The segmentation inputs were downsampled using linear interpolation in all axes, with no need for nearest neighbor interpolation on coarsely encoded inputs. Exact input shape and voxel sizes of the inputs can be found in **Supplementary Table 14.** We performed data augmentation using random three-dimensional affine and elastic transformations of the input volumes using the Multidimensional Image Augmentation library (see Code Availability section). Our deformation parameters are listed in **Supplementary Table 15**.

The network consists of six levels of three-dimensional convolutions organised into "blocks". A block consists of convolutions with 1x3x3 and 3x1x1 kernels with skip connections to a final concat operation where the outputs of all previous convolutions plus the input are stacked in the channel dimension (see **Extended Data Figure 10**). If the input has dimension $[z,\ y,\ x,\ c]$ and a block has n convolution with k channels each then the final output of a block would be $[z,\ y,\ x,\ c + n * k]$. Skip connections draw inspiration from Dense Blocks described in Huang et al. 2016[65], where each convolution receives the stacked outputs of all previous convolutions plus the input: the ith convolution receives an input of size $[z,\ y,\ x,\ c + (i - 1) * k]$ leading to an explosion of parameters but denser representations. However, we found the dense skip connections at every layer in the block to be dispensable in our case. Our blocks with single skip connections per layer save memory, use less parameters and still achieve the benefits from dense blocks such as better feature propagation and better gradients. The choice of 1x3x3 and 3x1x1 kernels is motivated by Xie et al. 2018[66] that found the factorised 1x3x3 and 3x1x1 saves memory and performs better than the full 3x3x3 convolution. A combination of 1x1x1 convolutions and three-dimensional max pooling operations were performed between consecutive levels to reduce the number of feature outputs from concatenated dense blocks. The output of the network is fed to a dense layer with a global pool average that outputs exAMD conversion predictions over future time windows ranging from 3 to 24 months as well as predictions for the auxiliary tasks of predicting additional diagnoses and referral decision (see Clinical referral and diagnosis network). For an exact description of the architecture see **Supplementary Table 16**.

The training loss is taken as the sum of the sigmoid cross entropy losses for the exAMD conversion and the disease components and the softmax cross entropy loss for the multi-class referral decision components. The following describes the loss function for the exAMD prediction model with multiple tasks. As described in the ensembling section, each model is independently trained and thus has weights that differ from those of the other models.

The loss function for each model task is given by the cross-entropy loss between the ground-truth labels y and the model prediction $f(x|\theta)$ given an input scan or segmentation map x:

$$H(\mathbf{y}, f(x|\theta)) = \sum_{k=1}^{K} -y_k \log(f_k(x|\theta))$$

where y_k is 1 for the correct class and 0 for the rest, and $f_k(x|\theta)$ is the model prediction for class k given the model weights θ.

As we describe in the paper, for the auxiliary diagnosis and clinical referral classification tasks that regularise the model, the same loss function is used with $y_k \in [0, 1]$ being the distillation labels, which are continuous due to being the prediction outputs[64] from the referral and diagnosis model (see Clinical referral and diagnosis network). Note that the exAMD conversion predictions and the disease classifications are binary classification tasks, and the referral classification is a multi-class task ($K = 4$).

The total loss function per input is defined as

$$\mathcal{L}_{total} = \overbrace{\sum_{t=3m}^{24m} \left[H(\mathbf{y}^{t}, f^{t})\right]}^{\text{Main loss}} + \overbrace{\sum_{disease} \left[H(\mathbf{y}^{disease}, f^{disease})\right] + H(\mathbf{y}^{ref}, f^{ref})}^{\text{Auxiliary loss}}.$$

with t being the time window for conversion predictions in months, and *disease* and *ref* being the auxiliary diagnosis and clinical referral classifications respectively.

Loss weighting was found to be crucial in training the models to favor the training loss in maximizing future-conversion performance. The number of post-conversion scans compared to pre-conversion constituted a 10:1 ratio which is reflected in the label distributions. Post-conversion scans were thus loss weighted 1/10 for auxiliary task to boost performance. Masking future conversion labels in post-conversion scans improved performance, as penalizing the model for incorrect future predictions once the event has occurred is illogical.

The hyperparameters were chosen based on performance on the validation set. Batch-norm, layer-norm, and dropout were ineffective in improving validation performance. Furthermore, minimal differences were found when using different model parameter settings for each input modality. Thus, the same hyperparameters were chosen for both the raw OCT input and segmentation input. The model was trained separately on each input without any parameter sharing. Training was performed with a batch size of 16 and a learning rate schedule starting with 0.0005 then set to 0.0005/8 after 60% of the total iterations, 0.0005/64 after 90%, and finally 0.0005/256 for the final 5% of training. Optimization was performed using Adam[67] with 1×10−5 weight decay, 0.9 β_1 and 0.98 β_2; learning rate warmup over 10000 iterations at a rate of 0.5. OCT training was run for 100000 iterations.

Cross-validation and ensembling

While hyperparameter tuning was carried out using a 20% validation set, cross-validation (CV) was used for final model ensembling due to the limited size of the dataset to prevent overfitting. The patients in the training and validation set were randomly partitioned into four folds at the patient level. Our final ensemble included model instances trained on each CV group (three folds used for training, one for validation). For each CV group, three instances of the exAMD prediction model with different random initializations were trained on three folds and evaluated on the validation fold. This was performed for both input types (raw OCT and segmentation map). The total number of trained models was 24, three randomly initialised instances for each of the two input modalities trained on each of the four CV groups. After training each model individually and freezing the weights, we ensembled all 24 models by taking the average over each of the models' outputs. Ensembling all 24 models resulted in the best performance with more instances giving insignificant improvements. At test time evaluation, we performed 10 instances of test-time augmentation (TTA) for each model using deformation parameters toned down from train-time deformation (**Supplementary Table 15**). We observed using TTA on the cross validation set improves performance but did not treat the number of TTA or any of the deformation parameters as a hyperparameter to avoid any subtle overfitting. In total we ensembled 240 different model outputs for each example in the test set to get the final system predictions. **Extended Data Figure 7** gives a diagram of our ensembling scheme.

Clinical analysis

The segmentation network comprises five instances of the segmentation model. For clinical analyses, we used the mean segmentation map, obtained by averaging the logits over the 5 instances. By equating each voxel to the volume it occupies, overall volumes of each tissue class can be derived. We further analysed the mean segmentation output using calculated volumes and computer vision algorithms to perform geometric categorisation of different tissue classes to derive clinically meaningful subgroups (**Supplementary Table 17**). Four different categories of subgroups were analysed: drusen volume, geographic atrophy (GA) presence, hyperreflective foci (HRF) presence, and features pathognomonic of exAMD that were present on the conversion scan (i.e. intraretinal fluid (IRF), subretinal fluid (SRF), subretinal hyperreflective material (SHRM), and fibrovascular pigment epithelium detachment (PED)). In addition, enface maps were produced to qualitatively analyse segmentation outputs.

Drusen staging

Drusen parameters such as diameter, height, area, and volume have been studied extensively and are known to correlate with exAMD conversion risk[30,31]. For this study, we explored conversion rates and system performance in ranges of drusen volume. To calculate the volume of drusen in the OCT scans, the drusenoid PED tissue class was isolated from each segmentation map for each scan in the test set. The distribution of drusen volume was stratified into 4 quartiles (0-25th percentile, 25-50th percentile, 50-75th percentile, 75-100th percentile). See **Supplementary Table 17** for further details.

Geographic atrophy presence

Geographic atrophy (GA) is identified by the attenuation of RPE tissue. GA is most easily visible on enface maps. To isolate areas of GA, a connected components algorithm was subsequently run on the pixels without RPE to find areas of atrophy. Each detected atrophy region was measured along each axis of the enface map to detect the largest diameter (major axis) of atrophy. GA was classified as present if the major axis had a diameter of ≥250 µm, as proposed by Sadda et al[68].

Hyperreflective foci presence

HRF are presented by relatively small hyperreflective regions within the neurosensory tissue on segmentation maps. We defined HRF as definitely present if a set of connected HRF voxels was ≥4 voxels, approximately equal to 5750 μm^3. This was determined using a connected components algorithm.

Conversion scan subgroups

For each converting fellow eye, we analysed the segmentation map on the visit determined to be where the eye converted. IRF, SRF, SHRM, and fibrovascular PED were classified as present if they had a volume greater than 5 voxels, approximately equal to 7200 μm^3. For the subgroup analysis, all scans prior to conversion in these eyes were analysed, stratified by the appearance of the conversion scan.

Enface maps

Given a 3-dimensional tissue segmentation we calculated an enface map per tissue by summing the number of voxels across the A-scan direction, generating a 2-dimensional map of tissue thicknesses across the scanned macula area. The result is a tissue heatmap across B and C scans. These can be plotted across time providing a useful summary of anatomical abnormalities across the full patient history.

Statistical analysis

To compute 95% confidence intervals for the true and false positive rates (i.e. sensitivity and 1-specificity), we used the Clopper-Pearson interval as implemented in the Python **statsmodels** library (v0.9.0). Kaplan-Meier survival curves were calculated using the Python **lifelines** library (v0.14.6). Inter-expert variability was calculated using Python sklearn.metrics library (v0.20.0). ROCAUC confidence intervals were computed via Bootstrap. P-values in **Supplementary Table 9**, **Extended Data Figure 3** were computed using two-sided permutation tests. P-values in **Supplementary Table 10** were computed with McNemar tests.

Reporting summary

Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Code availability

We make use of several open-source libraries to conduct our experiments, namely the machine learning framework TensorFlow (https://github.com/tensorflow/tensorflow) along with the TensorFlow library Sonnet (https://github.com/deepmind/sonnet) which provides implementations of individual model components [57]. For image augmentation we use the Multi-dimension image augmentation library previously open sourced by DeepMind (https://github.com/deepmind/multidim-image-augmentation). The model architecture is available open source (https://github.com/google-health/imaging-research). Other aspects of the experimental system make use of proprietary libraries and we are unable to publicly release this code. We detail the experiments and implementation details in the methods section and in the supplementary figures to allow for independent replication.

Data availability

The clinical data used for the training, validation and test sets were collected at Moorfields Eye Hospital NHS Foundation Trust and transferred to DeepMind in a de-identified format. Data were used with both local and national permissions. They are not publicly available and restrictions apply to their use. The data, or a test subset, may be available from Moorfields Eye Hospital NHS Foundation Trust subject to local and national ethical approvals. Moorfields Eye Hospital NHS Foundation Trust intends to make the raw data shared with DeepMind openly available to researchers as part of the Ryan Initiative for Macular Research (http://rimr.doheny.org/).

Methods-only References

49. Information Commissioner's Office. *Anonymisation: managing data protection risk code of practice*. (2015).
50. De Fauw, J. *et al.* Automated analysis of retinal imaging using machine learning techniques for computer vision. *F1000Res.* **5**, 1573 (2016).
51. Balaratnasingam, C., Yannuzzi, L.A., Curcio, C.A., Morgan, W.H., Querques, G., Capuano, V., Souied, E., Jung, J. and Freund, K.B.. Associations between retinal pigment epithelium and drusen volume changes during the lifecycle of large drusenoid pigment epithelial detachments. Investigative ophthalmology & visual science, 57(13), 5479-5489 (2016)
52. Balaratnasingam, C., Hoang, Q.V., Inoue, M., Curcio, C.A., Dolz-Marco, R., Yannuzzi, N.A., Dhrami-Gavazi, E., Yannuzzi, L.A. and Freund, K.B.. Clinical characteristics, choroidal neovascularization, and predictors of visual outcomes in acquired vitelliform lesions. American journal of ophthalmology, 172, 28-38 (2016)
53. Lek, J.J., Caruso, E., Baglin, E.K., Sharangan, P., Hodgson, L.A., Harper, C.A., Rosenfeld, P.J., Luu, C.D. and Guymer, R.H. Interpretation of subretinal fluid using OCT in intermediate age-related macular degeneration. Ophthalmology Retina, 2(8), 792-802 (2018)
54. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016* 424–432 (2016). doi:10.1007/978-3-319-46723-8_49
55. Abadi, M. *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. (2016).
56. Open sourcing Sonnet - a new library for constructing neural networks | DeepMind. *DeepMind* Available at: https://deepmind.com/blog/open-sourcing-sonnet/. (Accessed: 26th July 2019)

57. Buchlovsky, P. *et al.* TF-Replicator: Distributed Machine Learning for Researchers. (2019).

58. Curcio, C. A., Zanzottera, E. C., Ach, T., Balaratnasingam, C. & Freund, K. B. Activated Retinal Pigment Epithelium, an Optical Coherence Tomography Biomarker for Progression in Age-Related Macular Degeneration. *Invest. Ophthalmol. Vis. Sci.* **58**, BIO211–BIO226 (2017).

59. Christenbury, J.G., Folgar, F.A., O'Connell, R.V., Chiu, S.J., Farsiu, S., Toth, C.A. and Age-related Eye Disease Study 2 Ancillary Spectral Domain Optical Coherence Tomography Study Group. Progression of intermediate age-related macular degeneration with proliferation and inner retinal migration of hyperreflective foci. *Ophthalmology*, **120(5)**, 1038-1045 (2013).

60. Folgar, F. A. *et al.* Spatial correlation between hyperpigmentary changes on color fundus photography and hyperreflective foci on SDOCT in intermediate AMD. *Invest. Ophthalmol. Vis. Sci.* **53**, 4626–4633 (2012).

61. Age-Related Eye Disease Study Research Group. Risk factors associated with age-related macular degeneration. A case-control study in the age-related eye disease study: Age-Related Eye Disease Study Report Number 3. *Ophthalmology* **107**, 2224–2232 (2000).

62. Klein, R., Klein, B. E., Jensen, S. C. & Meuer, S. M. The five-year incidence and progression of age-related maculopathy: the Beaver Dam Eye Study. *Ophthalmology* **104**, 7–21 (1997).

63. Bressler, S. B., Maguire, M. G., Bressler, N. M. & Fine, S. L. Relationship of drusen and abnormalities of the retinal pigment epithelium to the prognosis of neovascular macular degeneration. The Macular Photocoagulation Study Group. *Arch. Ophthalmol.* **108**, 1442–1447 (1990).

64. Hinton, G., Vinyals, O. & Dean, J. Distilling the Knowledge in a Neural Network. (2015).

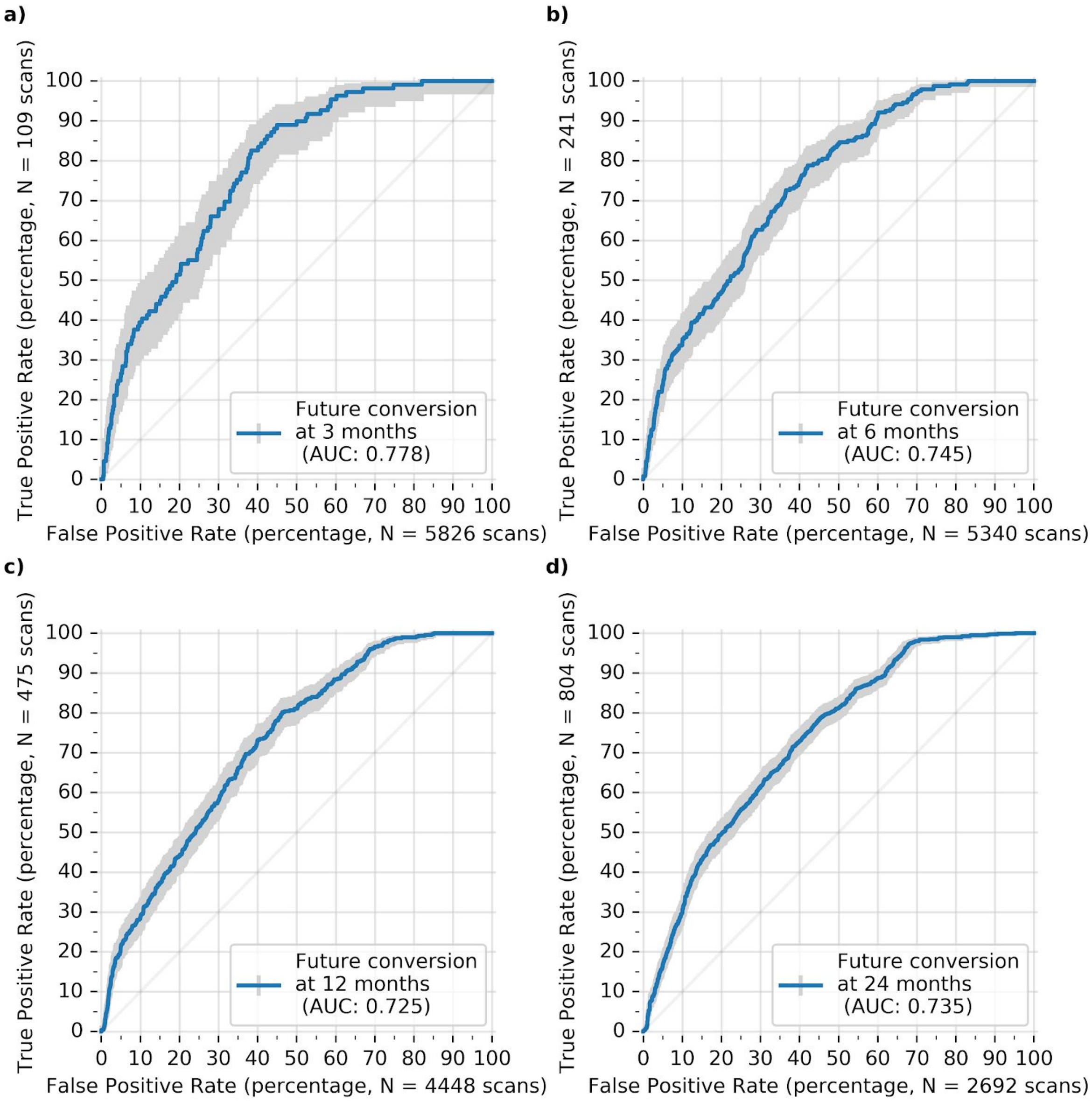
65. Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. Densely connected convolutional networks. *In Proceedings of the IEEE conference on computer vision and pattern recognition.* 4700-4708 (2017).

66. Xie, S., Sun, C., Huang, J., Tu, Z. and Murphy, K.. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In Proceedings of the European Conference on Computer Vision (ECCV), 305-321 (2018).

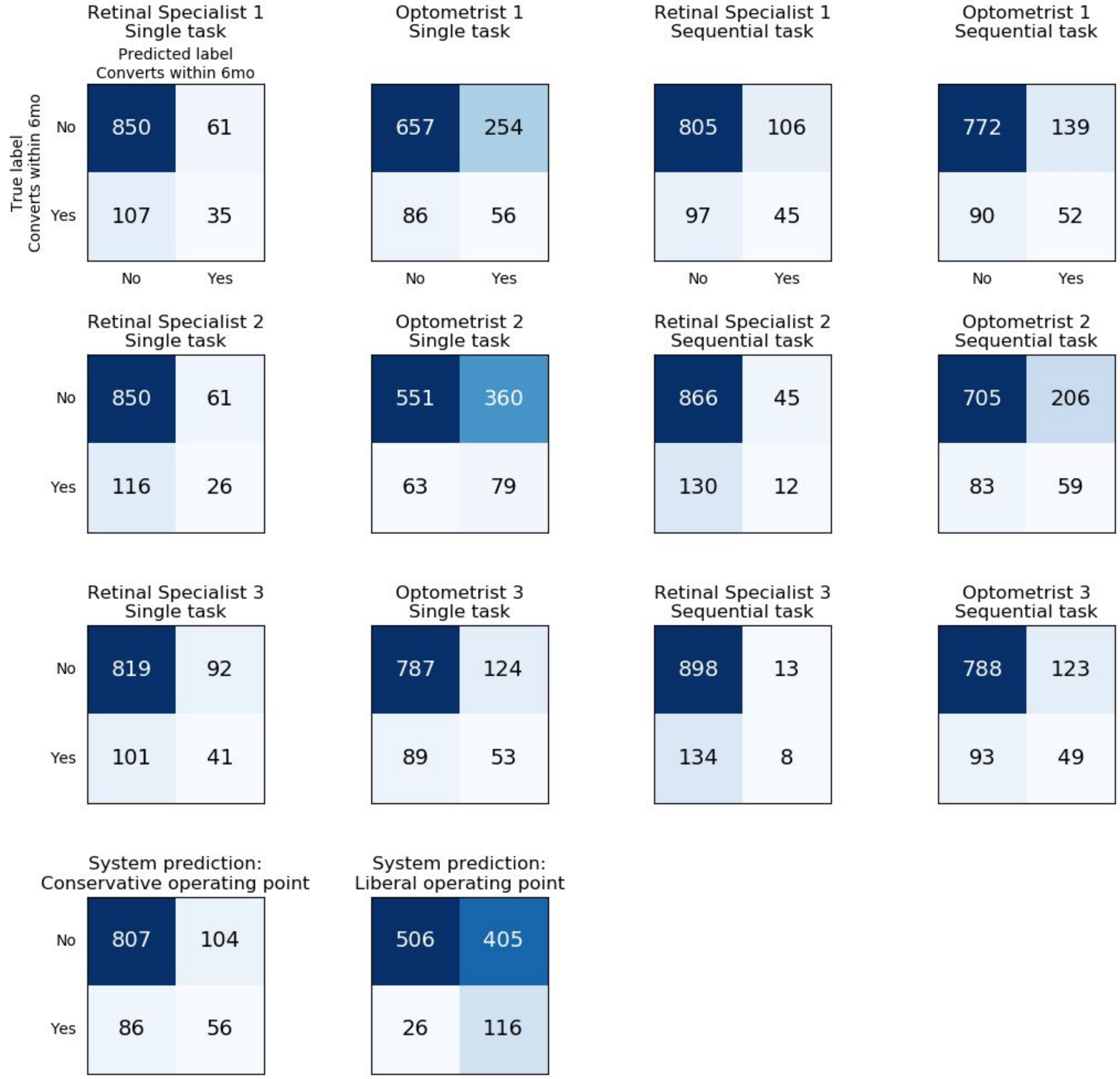
67. Kingma, D.P. and Ba, J.. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

68. Sadda, S. R. *et al.* Consensus Definition for Atrophy Associated with Age-Related Macular Degeneration on OCT: Classification of Atrophy Report 3. *Ophthalmology* **125**, 537–548 (2018).

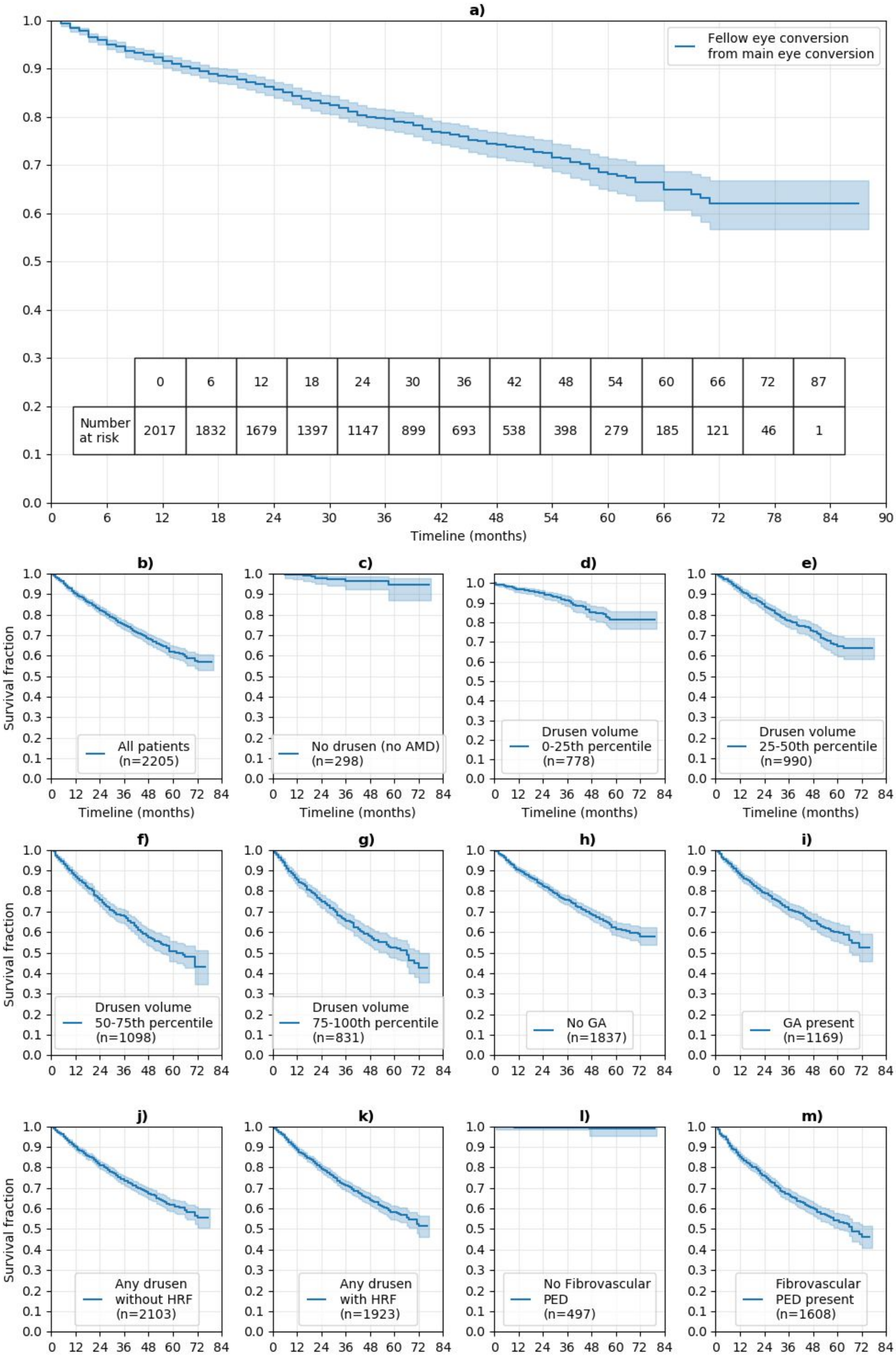
Extended Data



Extended Data Figure 3 | ROC curves for various time windows. ROC curves for model predictions over time windows of **a**) 3 months, **b**) 6 months, **c**) 12 months, and **d**) 24 months. Note that the difference in AUC between 12 and 24 month predictions is not statistically significant (p-value=0.54, two-sided permutation test).

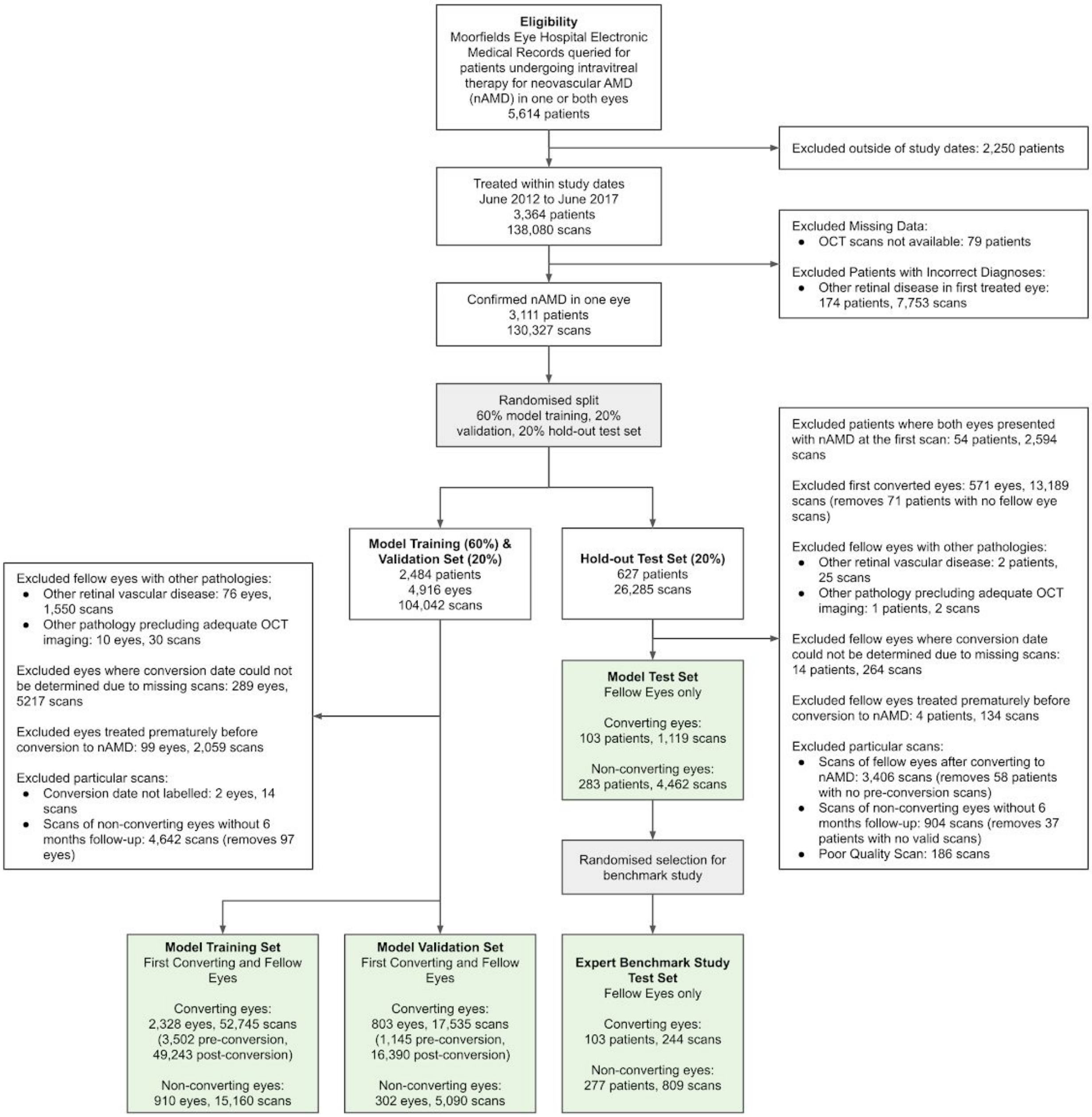


Extended Data Figure 4 | Confusion matrices per expert and task. Confusion matrices for the prediction decision for all 6 experts for the single scan and sequential scan tasks, and for the system at two chosen operating points**.** n=1053 trials (380 unique patients).

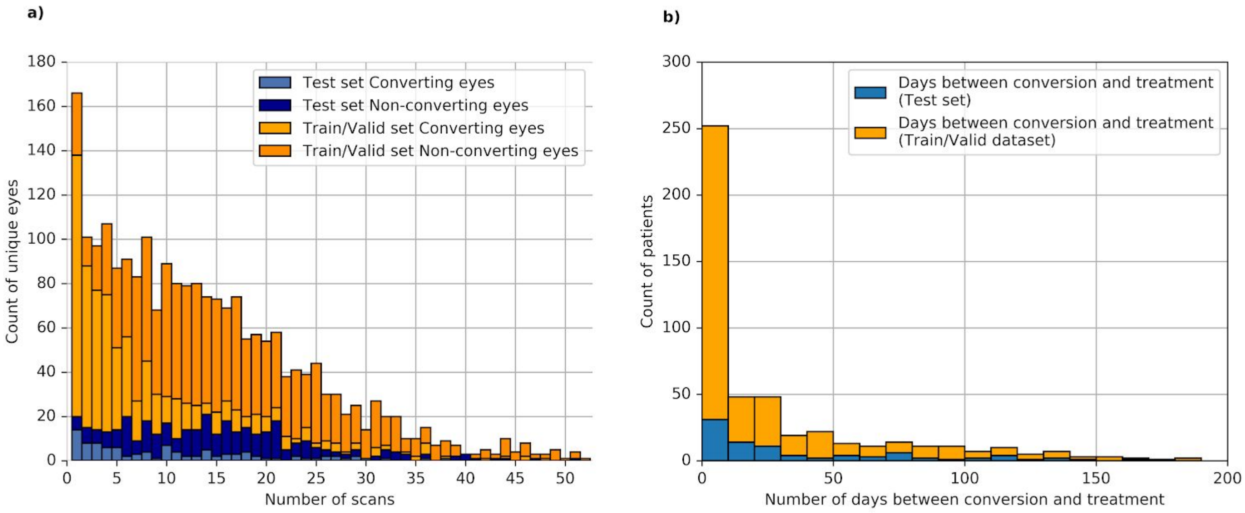


	0	6	12	18	24	30	36	42	48	54	60	66	72	87
Number at risk	2017	1832	1679	1397	1147	899	693	538	398	279	185	121	46	1

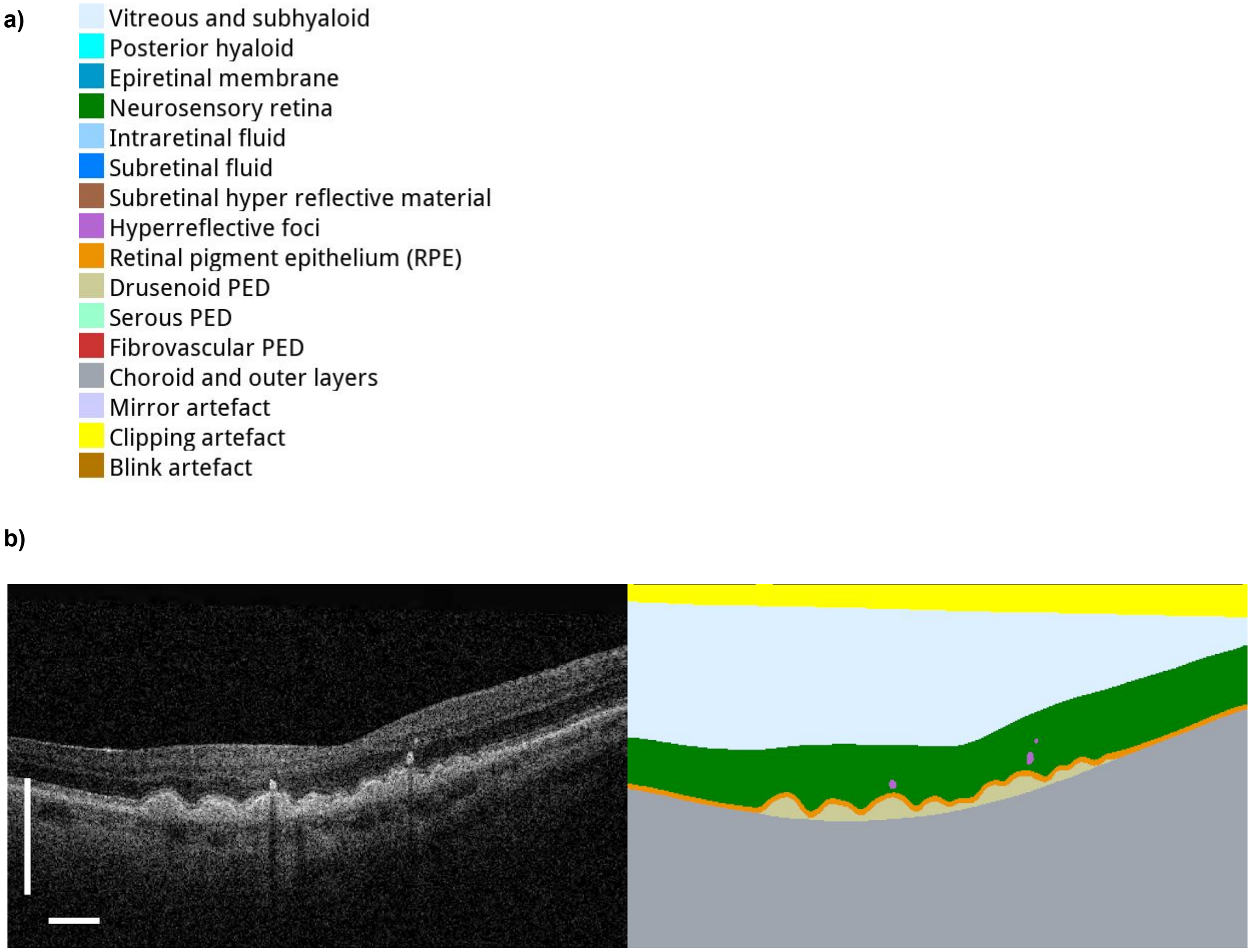
Extended Data Figure 7 | Kaplan-Meier survival curves for full dataset and subgroups stratified by drusen stage and presence of HRF. a) A Kaplan-Meier survival curve for fellow eye conversion to exAMD from baseline (defined as the first presentation of first eye conversion) in number of months, showing a little over 40% of patients converted during over 6 years of available follow up. The table shows number of eyes remaining at risk per month. **b)** The same plot for comparison with following plots. **c-g**) Plots for varying amounts of drusen, showing increasing numbers of patient convert as drusen volume increases. Drusen size categories are calculated as quartiles. The same plots are shown for patients **h**) without and i) with geographic atrophy (GA), those **j**) without and **k**) with hyper-reflective foci (HRF), and those **l**) without and **m**) with fibrovascular pigment epithelial detachment (PED). In all plots the timeline is with reference to the first incidence of the feature in the eye.


Eligibility
Moorfields Eye Hospital Electronic Medical Records queried for patients undergoing intravitreal therapy for neovascular AMD (nAMD) in one or both eyes
5,614 patients
Excluded outside of study dates: 2,250 patients
Treated within study dates
June 2012 to June 2017
3,364 patients
138,080 scans
Excluded Missing Data:
• OCT scans not available: 79 patients
Excluded Patients with Incorrect Diagnoses:
• Other retinal disease in first treated eye: 174 patients, 7,753 scans
Confirmed nAMD in one eye
3,111 patients
130,327 scans
Randomised split
60% model training, 20% validation, 20% hold-out test set
Excluded patients where both eyes presented with nAMD at the first scan: 54 patients, 2,594 scans
Excluded first converted eyes: 571 eyes, 13,189 scans (removes 71 patients with no fellow eye scans)
Excluded fellow eyes with other pathologies:
• Other retinal vascular disease: 2 patients, 25 scans
• Other pathology precluding adequate OCT imaging: 1 patients, 2 scans
Excluded fellow eyes where conversion date could not be determined due to missing scans: 14 patients, 264 scans
Excluded fellow eyes treated prematurely before conversion to nAMD: 4 patients, 134 scans
Excluded particular scans:
• Scans of fellow eyes after converting to nAMD: 3,406 scans (removes 58 patients with no pre-conversion scans)
• Scans of non-converting eyes without 6 months follow-up: 904 scans (removes 37 patients with no valid scans)
• Poor Quality Scan: 186 scans
Model Training (60%) & Validation Set (20%)
2,484 patients
4,916 eyes
104,042 scans
Hold-out Test Set (20%)
627 patients
26,285 scans
Excluded fellow eyes with other pathologies:
• Other retinal vascular disease: 76 eyes, 1,550 scans
• Other pathology precluding adequate OCT imaging: 10 eyes, 30 scans
Excluded eyes where conversion date could not be determined due to missing scans: 289 eyes, 5217 scans
Excluded eyes treated prematurely before conversion to nAMD: 99 eyes, 2,059 scans
Excluded particular scans:
• Conversion date not labelled: 2 eyes, 14 scans
• Scans of non-converting eyes without 6 months follow-up: 4,642 scans (removes 97 eyes)
Model Test Set
Fellow Eyes only
Converting eyes:
103 patients, 1,119 scans
Non-converting eyes:
283 patients, 4,462 scans
Randomised selection for benchmark study
Model Training Set
First Converting and Fellow Eyes
Converting eyes:
2,328 eyes, 52,745 scans (3,502 pre-conversion, 49,243 post-conversion)
Non-converting eyes:
910 eyes, 15,160 scans
Model Validation Set
First Converting and Fellow Eyes
Converting eyes:
803 eyes, 17,535 scans (1,145 pre-conversion, 16,390 post-conversion)
Non-converting eyes:
302 eyes, 5,090 scans
Expert Benchmark Study Test Set
Fellow Eyes only
Converting eyes:
103 patients, 244 scans
Non-converting eyes:
277 patients, 809 scans

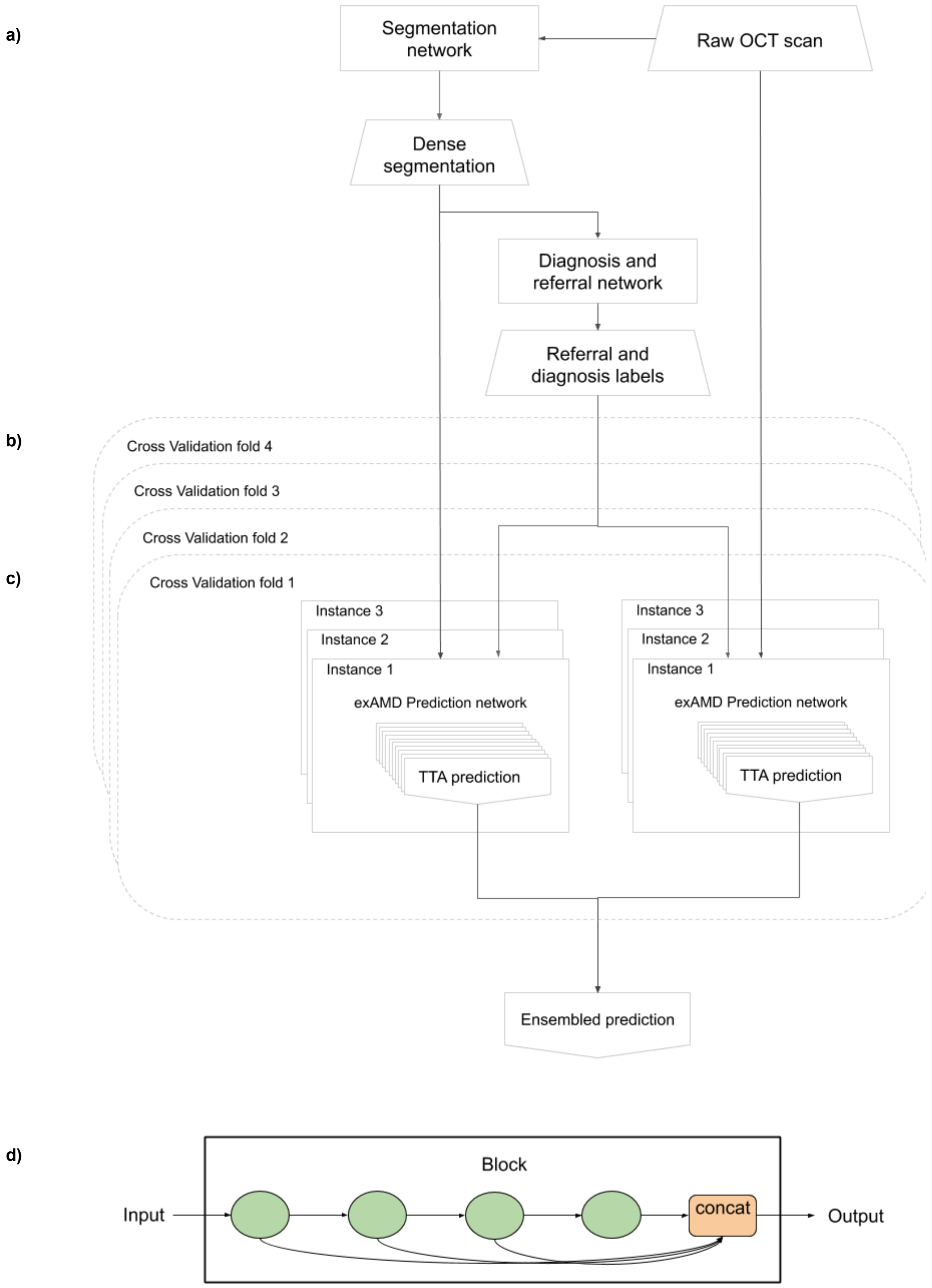
Extended Data Figure 8 | Consort diagram. Data labelling of the Moorfields Eye Hospital AMD dataset. Manual opt outs before data transfer are not included as none of the patients who manually opted out had digital OCT within the study dates.



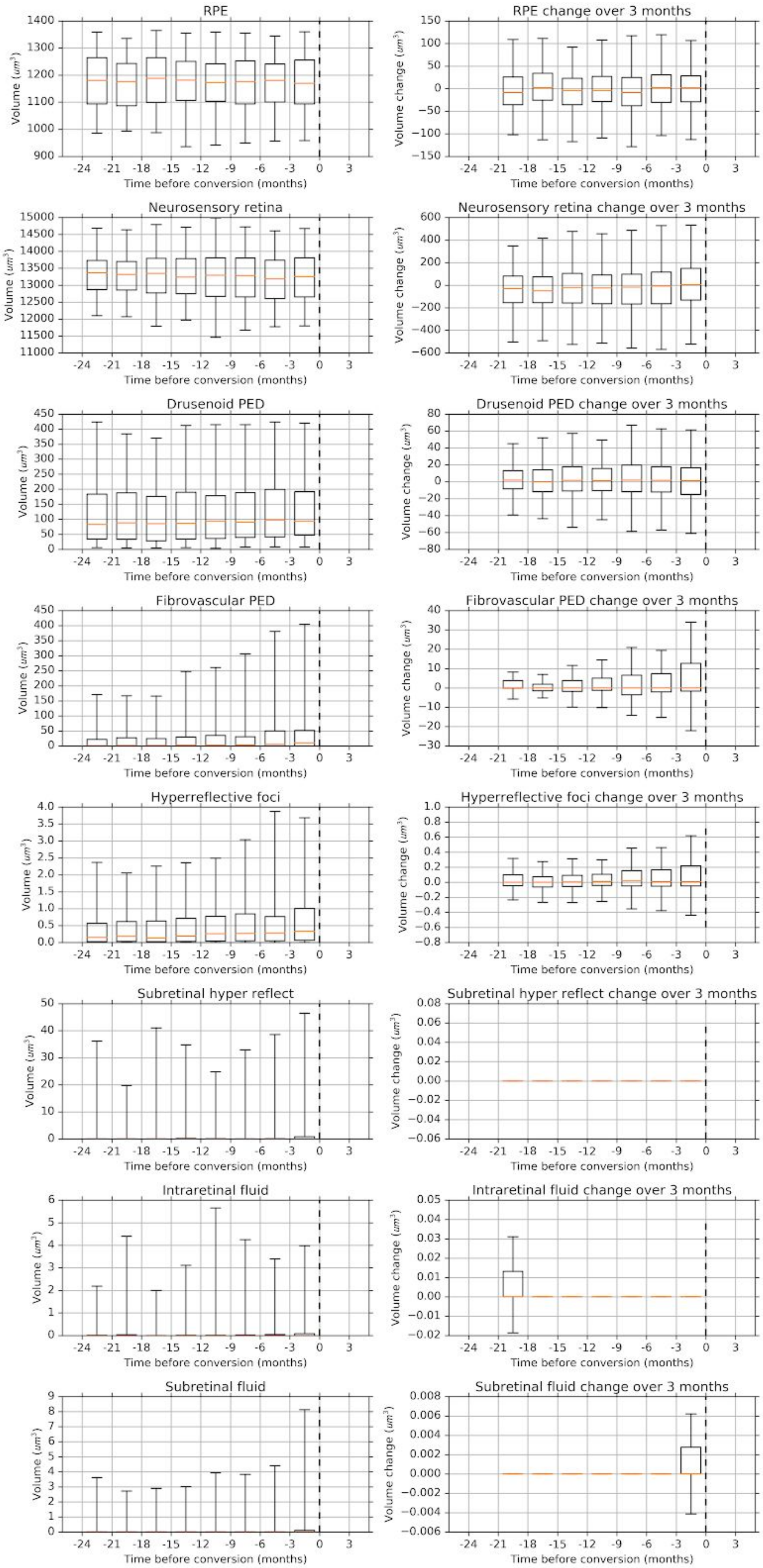
Extended Data Figure 2 | Dataset statistics. (a) Histogram of scans per unique eye (pre-conversion) in training/validation set and test set. **(b)** Histogram of difference between conversion and injection date for fellow eyes in training and test set (n=537, 214 have matching dates).



Extended Data Figure 9 | Segmentation colour key and new hyperreflective foci class. a) Colour key for 13 tissues and 3 artefacts segmented by the network. **b)** Left: Raw OCT input to the segmentation network. Right: Output of the retrained segmentation network. Three hyperreflective foci apparent in the intraretinal layers were successfully segmented in this B-scan (purple).


a)
Segmentation network
Raw OCT scan
Dense segmentation
Diagnosis and referral network
Referral and diagnosis labels
b)
Cross Validation fold 4
Cross Validation fold 3
Cross Validation fold 2
c)
Cross Validation fold 1
Instance 3
Instance 2
Instance 1
exAMD Prediction network
TTA prediction
Instance 3
Instance 2
Instance 1
exAMD Prediction network
TTA prediction
Ensembled prediction
d)
Block
Input
concat
Output

Extended Data Figure 7 | **Deep learning system diagram.** Flow chart of the deep learning system including ensembling and TTA. Model inputs are shaped as trapezoids. Deep learning networks are shaped as rectangles. Model outputs are shaped as pointed rectangles. **a**) The segmentation network takes a raw OCT scan as input to generate a dense segmentation of the OCT which is then fed into a Diagnosis and referral network to obtain auxiliary task referral and diagnosis labels. **b**) The auxiliary labels along with either the raw OCT scan or dense segmentation are inputted into each exAMD prediction network across each cross validation fold group. Although the arrows apply to one fold group and instance, they generalise across all fold groups and instances. **c**) Ten TTA predictions are obtained from each instance. All TTA predictions are combined via averaging to obtain the final ensembled prediction. **d)** Architecture of a single block in our network. Green circles are convolution layers applied sequentially to the input of the previous layer. Each convolution has stride 1 and uses ReLU activation. Four convolutions are shown for demonstrative purposes but the number of convolutions and the kernels used for each will differ between blocks. Each convolution has a skip connection to the last orange node which concatenates all the intermediate and final activations along the channel dimension as the output.


RPE
Volume (um^3)
1400
1300
1200
1100
1000
900
-24 -21 -18 -15 -12 -9 -6 -3 0 3
Time before conversion (months)
RPE change over 3 months
Volume change (um^3)
150
100
50
0
-50
-100
-150
Neurosensory retina
15000
14500
14000
13500
13000
12500
12000
11500
11000
Neurosensory retina change over 3 months
600
400
200
0
-200
-400
-600
Drusenoid PED
450
400
350
300
250
200
150
100
50
0
Drusenoid PED change over 3 months
80
60
40
20
0
-20
-40
-60
-80
Fibrovascular PED
Fibrovascular PED change over 3 months
40
30
20
10
0
-10
-20
-30
Hyperreflective foci
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Hyperreflective foci change over 3 months
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
Subretinal hyper reflect
50
40
30
20
10
0
Subretinal hyper reflect change over 3 months
0.08
0.06
0.04
0.02
0.00
-0.02
-0.04
-0.06
Intraretinal fluid
6
5
4
3
2
1
0
Intraretinal fluid change over 3 months
0.05
0.04
0.03
0.02
0.01
0.00
-0.01
-0.02
Subretinal fluid
9
8
7
6
5
4
3
2
1
0
Subretinal fluid change over 3 months
0.008
0.006
0.004
0.002
0.000
-0.002
-0.004
-0.006

Extended Data Figure 6 | Aggregate volumes and volume change per 3 months before conversion for major ocular structures and abnormal tissues in patients who converted (n=549 unique patients). The box extends from the lower to upper quartile values of the data, with an orange line at the median. The whiskers show the 5th & 95th percentiles. For the left column, the statistics are calculated across patients, where patients with multiple scans per quarter are volumes averaged across these scans. For the right column the statistics for volume change over 3 months were calculated on the difference for each patient between the mean volume for that quarter against the previous quarter. Volumes are calculated using the whole 2.3*6*6mm OCT volume.

Extended Data Figure 1 | Summary statistics and patient demographic data. A breakdown of training (60%), validation (20%) and test (20%) datasets by unique patients and unique scans.

Patients and demographics		**Training**	**Validation**	**Test**	**Total**
Unique patients		2,019	669	676	**3,364**
Unique patients with scans		1,964	658	662	**3,284**
Unique patients after exclusions applied		1,795	614	386	**2,795**
Age at first eye exAMD presentation (baseline) - mean (SD)		79.2 (8.4)	79.9 (8.6)	78.8 (8.4)	**79.3 (8.5)**
Age at fellow eye conversion - mean (SD)		81.6 (7.3)	82.0 (7.8)	82.3 (7.3)	**81.8 (7.4)**
Ethnicity	White British and Irish	1000	322	206	**1,528**
	Asian	179	60	42	**281**
	Black	33	9	10	**52**
	Other	399	160	89	**648**
	Unknown	184	63	39	**286**
Sex	Female	1,073	400	240	**1,713**
	Male	720	213	146	**1,079**
	Unknown	2	1	0	3
Visual acuity at first eye exAMD presentation (baseline) - mean (SD)		55.3 (16.8), n=757	56.0 (17.2), n=257	55.5 (16.7), n=282	**55.5 (16.8), n=1296**
Visual acuity of fellow eye at baseline - mean (SD)		69.6 (17.3), n=671	68.9 (17.3), n=217	70.2 (17.8), n=235	**69.6 (17.4), n=1123**
Visual acuity at fellow eye conversion - mean (SD)		62.1 (16.8), n=347	62.1 (16.9), n=117	64.4 (17.3), n=108	**62.5 (16.9), n=572**
Eyes where visual acuity of fellow eyes at conversion: ≥71 letters / ≤70 letters / unknown VA		105 / 242 / 160	39 / 78 / 49	43 / 65 / 65	**187 / 385 / 274, n=846**
Number of days between fellow eye conversion and treatment - mean (SD)		64.0 (149.8), n=329	66.3 (147.4), n=111	66.4 (165.1), n=97	**64.9 (151.9), n=537**
Number of days between fellow eye conversion and treatment - median (IQR)		9.0 (0.0-50.0)	12.0 (0.0-59.5)	21.0 (0.0-77.0)	**13.0 (0.0-60.0)**
Number of days between fellow eye conversion and treatment where visual acuity at conversion ≥71 letters - mean (SD)		128.3 (198.5), n=97	141.2 (221.8), n=35	136.2 (253.5), n=36	**132.7 (214.9), n=168**
Number of days between fellow eye conversion and treatment where visual acuity at conversion ≤70 letters - mean (SD)		37.1 (114.0), n=232	31.8 (75.9), n=76	25.2 (38.6), n=61	**34.0 (98.0), n=369**
Scans					
Unique sequential scans	Pre-conversion to exAMD	18,662	6,235	5,581	**30,478**
	Post-conversion to exAMD	49,243	16,390	0	**65,633**
	Total scans	67,905	22,625	5,581	**96,111**

Extended Data Figure 5 | **Clinical expert metrics on the benchmark study. (a)** Metrics for each expert for the single scan and sequential scan tasks. Intra-observer agreement was assessed using Fleiss' Kappa. (PPV: positive predictive value, NPV: negative predictive value). **(b)** Agreement between the clinical experts for the single and sequential tasks, measured using Fleiss' Kappa. N=1053 for both single and sequential task.

a)

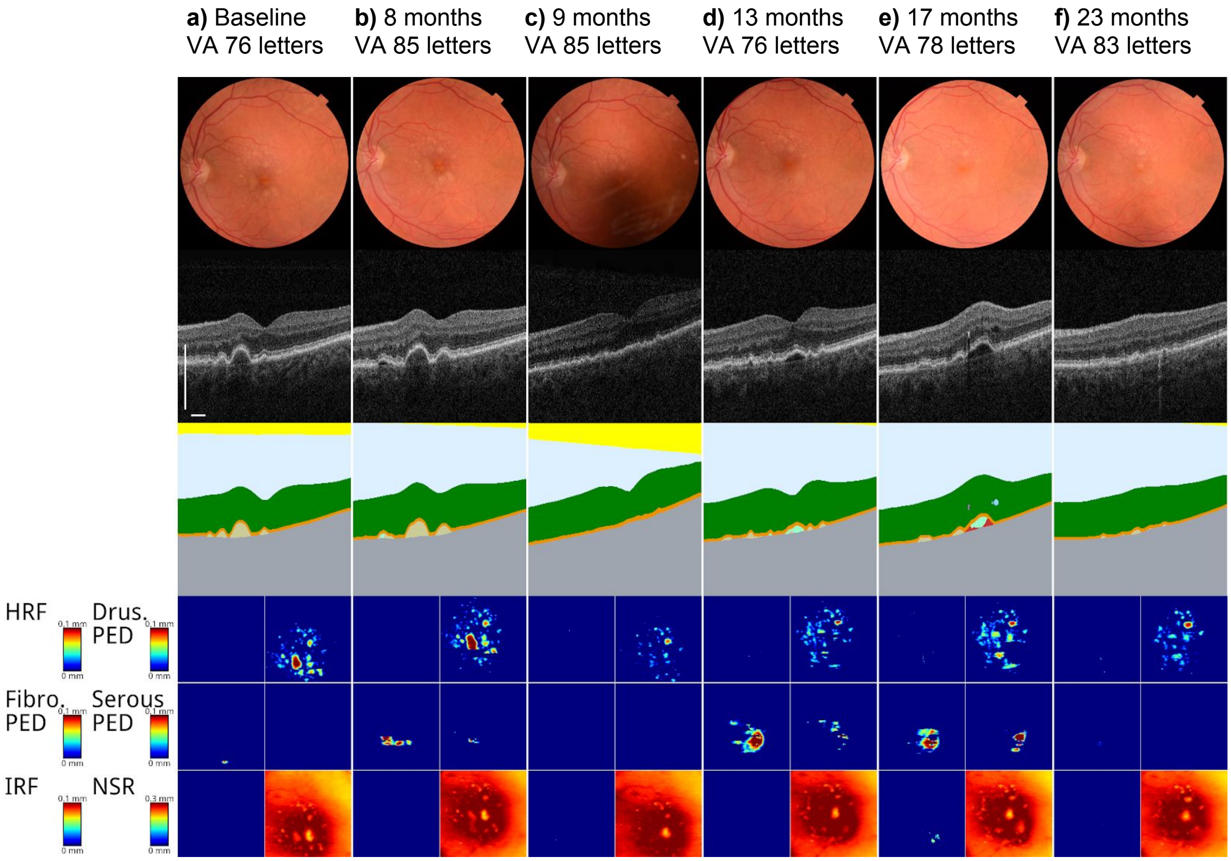
Expert	Task	Sensitivity (%)	Specificity (%)	Accuracy (%)	PPV (%)	NPV (%)	Kappa	Kappa 95% CI
Retinal Specialist 1	Single	24.6	93.3	84.0	36.5	88.8	0.32	0.239, 0.404
	Sequential	31.7	88.4	80.7	29.8	89.2		
Retinal Specialist 2	Single	18.3	93.3	83.2	29.9	88.0	0.316	0.212, 0.421
	Sequential	8.50	95.1	83.4	21.1	86.9		
Retinal Specialist 3	Single	28.9	89.9	81.7	30.8	89.0	0.180	0.099, 0.260
	Sequential	5.60	98.6	86.0	38.1	87.0		
Optometrist 1	Single	39.4	72.1	67.7	18.1	88.4	0.287	0.224, 0.350
	Sequential	36.6	84.7	78.3	27.2	89.6		
Optometrist 2	Single	55.6	60.5	59.8	18.0	89.7	0.292	0.236, 0.348
	Sequential	41.5	77.4	72.6	22.3	89.5		
Optometrist 3	Single	37.3	86.4	79.8	29.9	89.8	0.523	0.453, 0.592
	Sequential	34.5	86.5	79.5	28.5	89.4		

b)

Single Task	**Kappa**	**95% CI**
Between all Retinal Specialists (n=3)	0.335	0.267, 0.399
Retinal Specialist 1 & Retinal Specialist 2	0.384	0.288, 0.479
Retinal Specialist 1 & Retinal Specialist 3	0.302	0.216, 0.387
Retinal Specialist 2 & Retinal Specialist 3	0.333	0.247, 0.420
Between all Optometrists (n=3)	0.258	0.215, 0.304
Optometrist 1 & Optometrist 2	0.272	0.215, 0.330
Optometrist 1 & Optometrist 3	0.323	0.261, 0.386
Optometrist 2 & Optometrist 3	0.240	0.188, 0.291
Between all Retinal Specialists & Optometrists (n=6)	0.243	0.210, 0.280
Sequential Task	**Kappa**	**95% CI**
Between all Retinal Specialists (n=3)	0.143	0.082, 0.214
Retinal Specialist 1 & Retinal Specialist 2	0.165	0.088, 0.242
Retinal Specialist 1 & Retinal Specialist 3	0.132	0.063, 0.202
Retinal Specialist 2 & Retinal Specialist 3	0.208	0.083, 0.332
Between all Optometrists (n=3)	0.306	0.254, 0.652
Optometrist 1 & Optometrist 2	0.300	0.234, 0.366
Optometrist 1 & Optometrist 3	0.291	0.219, 0.364
Optometrist 2 & Optometrist 3	0.335	0.269, 0.401
Between all Retinal Specialists & Optometrists (n=6)	0.204	0.168, 0.242

Supplementary Information for *Predicting exudative conversion in age related macular degeneration using deep learning*

Supplementary Figures



a) Baseline VA 76 letters

b) 8 months VA 85 letters

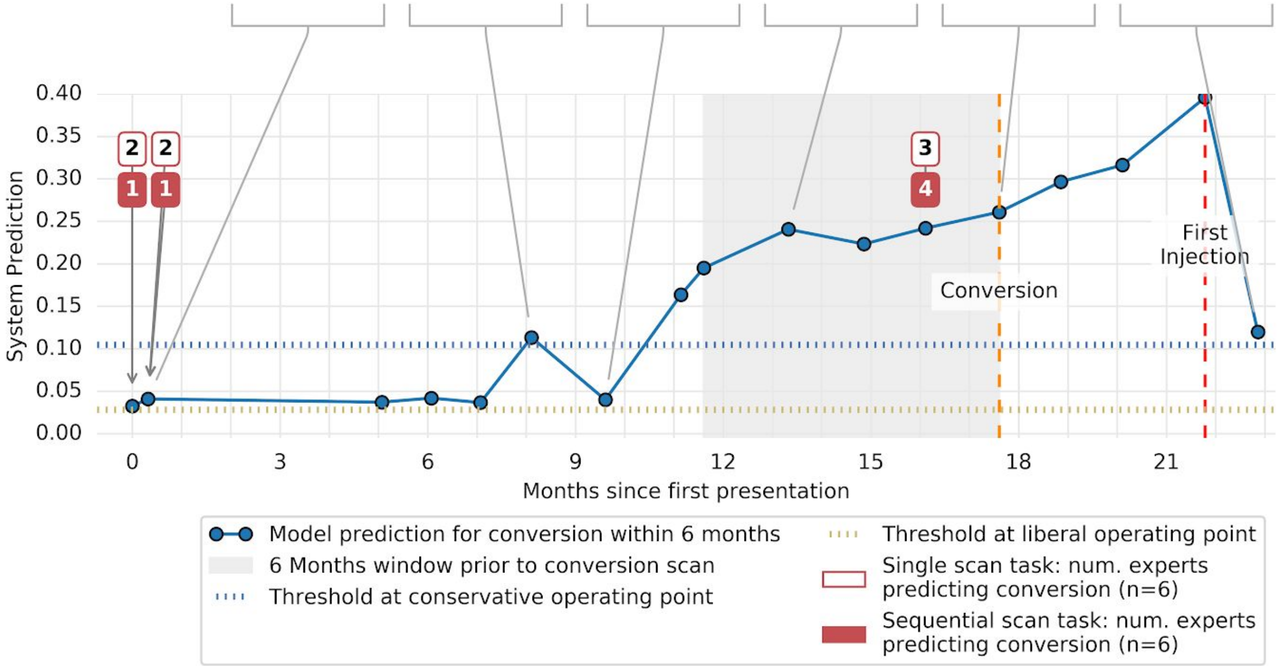
c) 9 months VA 85 letters

d) 13 months VA 76 letters

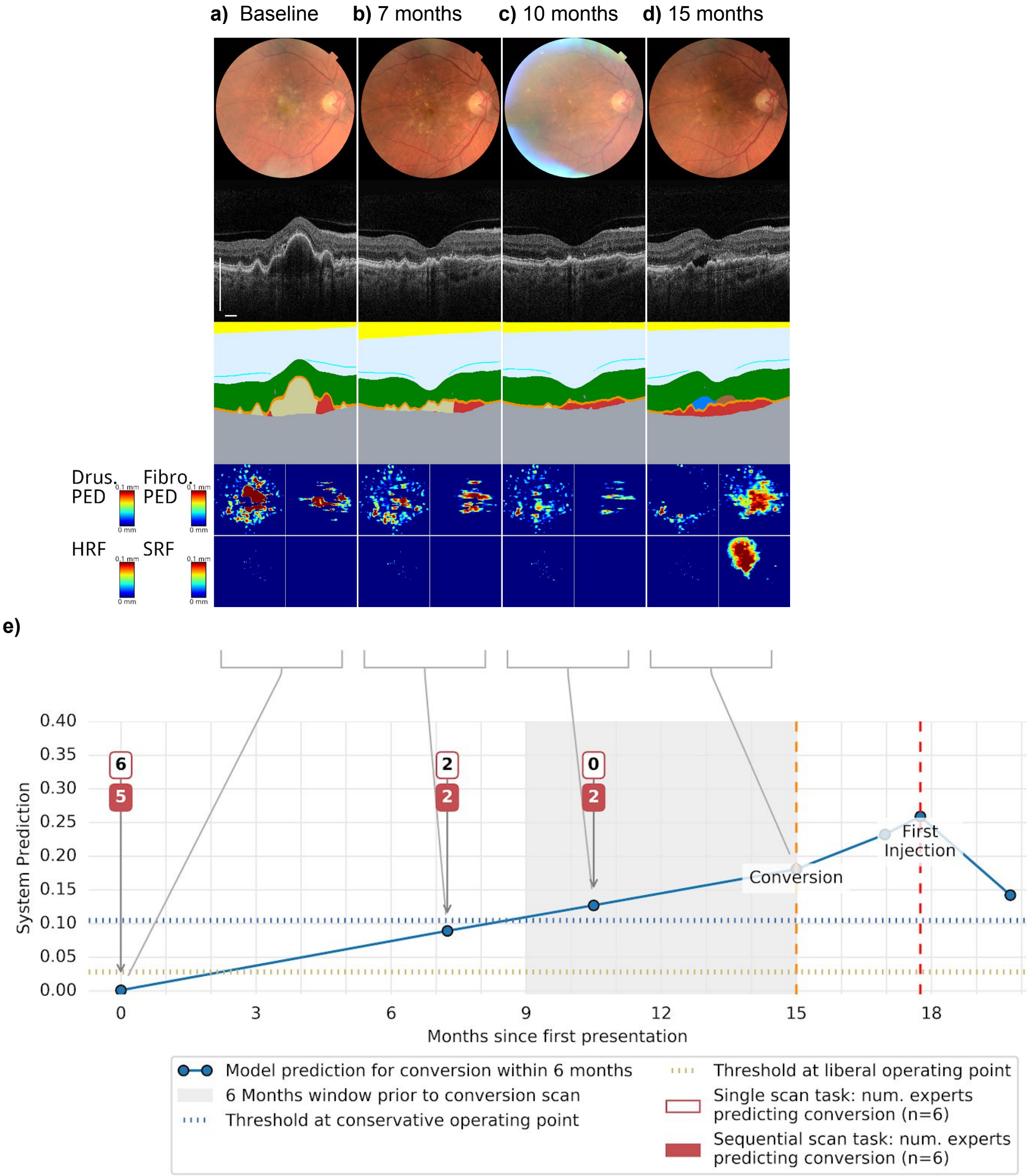
e) 17 months VA 78 letters

f) 23 months VA 83 letters

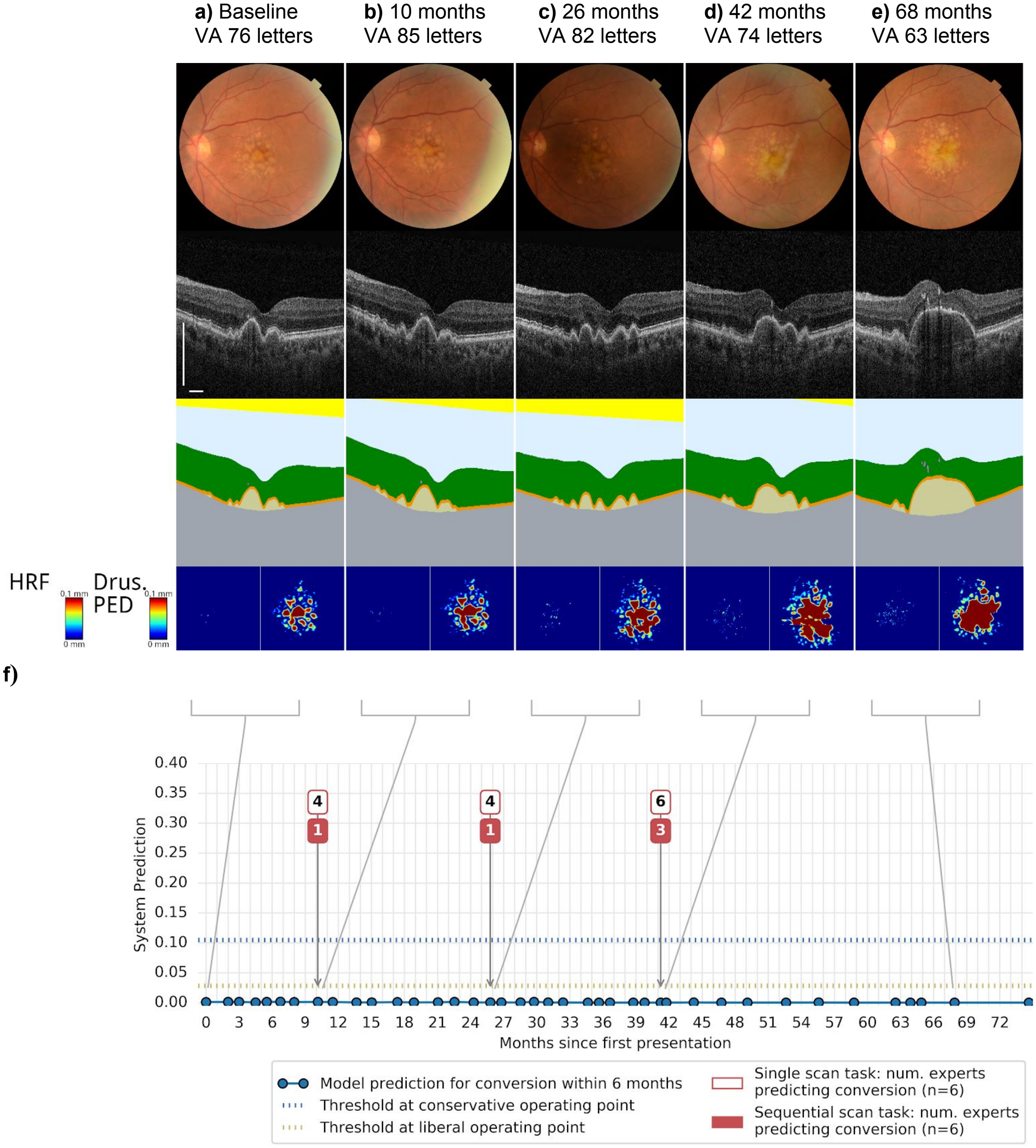
g)



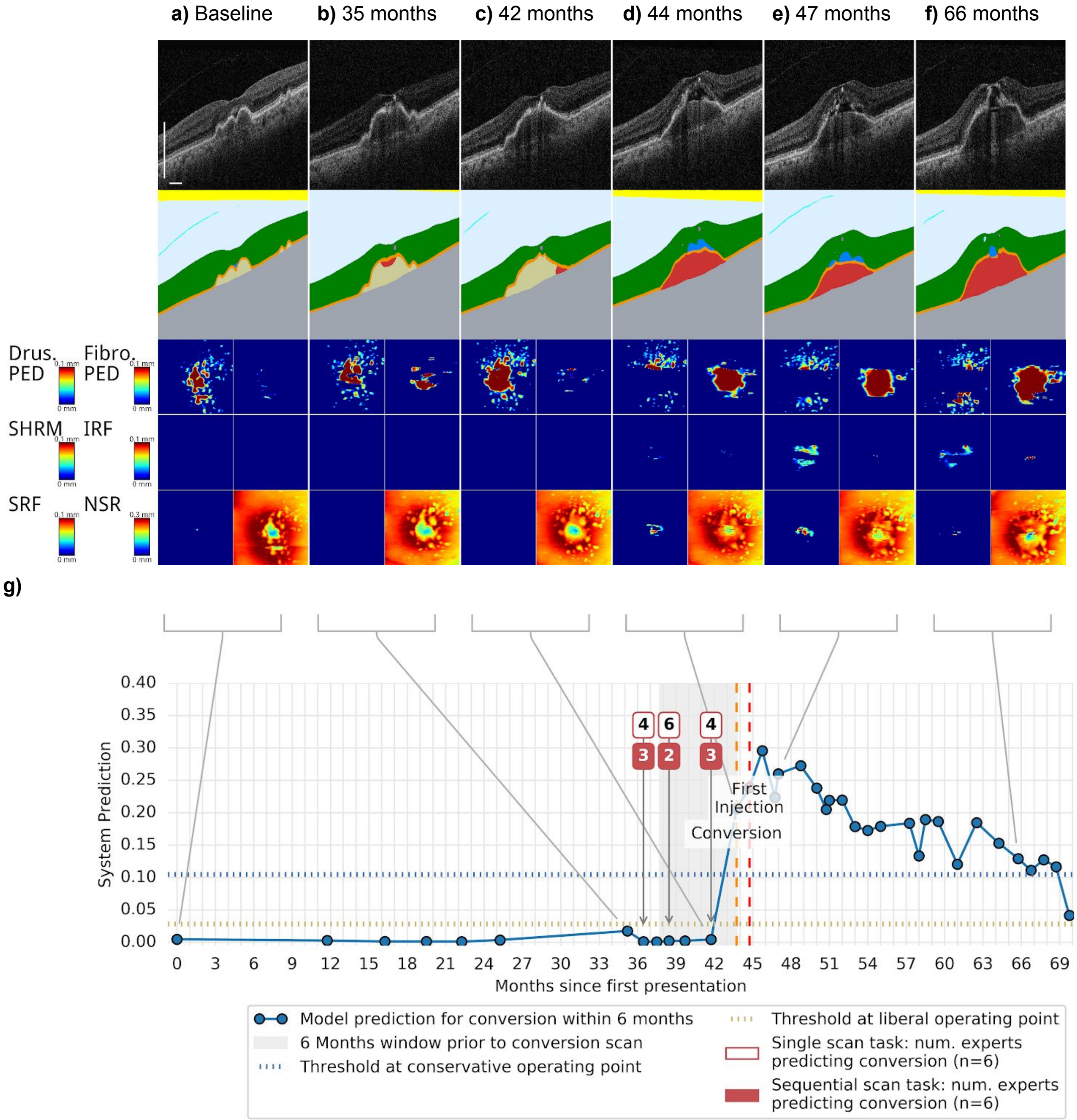
Supplementary Figure 1 | A true positive success case of an 81 year old white female presenting with exAMD in her right eye. Her fellow eye (left eye) shown in the figure is routinely scanned at every visit. **a)** The fellow eye presents with large drusen of highest density inferior-nasal to the fovea. This is clearly seen in the enface map. **b)** The drusen increases slightly in volume within 7 months. **c)** 9 months after baseline, the drusen has significantly and rapidly regressed in this area. **d)** Over time, a new shallow area of PED appears. Here the segmentation model predicts this has serous and fibrovascular elements - both suggestive of choroidal neovascularisation. However, no fluid is observed at this point and therefore observation continues. **e)** 4 months after the previous scan, cysts of intraretinal fluid are observed. Here, the eye is labelled and diagnosed with exAMD. As the vision is still as high as 78 letters, this particular patient does not receive treatment until 4 months later when the vision drops to 65 letters. **f)** After treatment, visual acuity improves and both intraretinal fluid and fibrovascular PED are markedly reduced. **g)** Risk prediction graph showing risk of imminent conversion rising above both operating point thresholds from slightly before the 6 months window prior to conversion. Note that there are early false positives ahead of conversion at both liberal and conservative operating points. (Enface map abbreviations, HRF: hyperreflective foci; Drus. PED: drusenoid PED; Fibro. PED: fibrovascular PED; IRF: intraretinal fluid; NSR: neurosensory retina). This and all subsequent success or failure cases were selected from the models most confident predictions.


a) Baseline
b) 7 months
c) 10 months
d) 15 months
Drus. PED
Fibro. PED
HRF
SRF
0.1 mm
0 mm
e)
6
5
2
2
0
2
First Injection
Conversion
System Prediction
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
3
6
9
12
15
18
Months since first presentation
Model prediction for conversion within 6 months
6 Months window prior to conversion scan
Threshold at conservative operating point
Threshold at liberal operating point
Single scan task: num. experts predicting conversion (n=6)
Sequential scan task: num. experts predicting conversion (n=6)

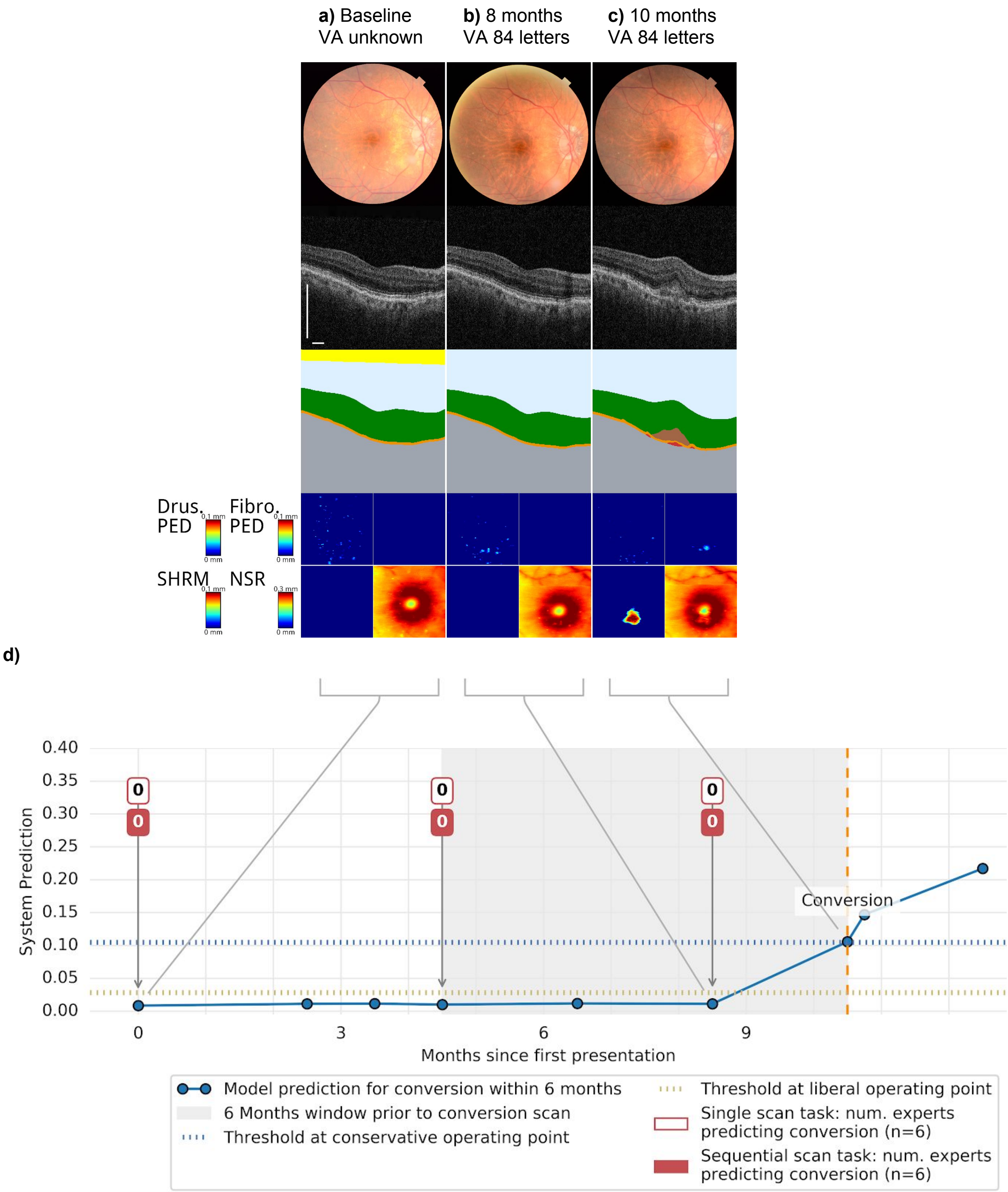
Supplementary Figure 2 | **A true positive success case of a 75 year old white British male presenting with exAMD in his left eye**. **a)** The right fellow eye presented with a large central drusenoid PED. Here the model segments the majority of this tissue as drusen, with some elements of fibrovascular. No fluid is present. **b)** At the next scan available in 7 months, the drusenoid PED has collapsed (drusen regression), leaving behind an irregular PED. **c)** 3 months later, the scan appears quite similar with both elements of fibrovascular and drusenoid PED being segmented. **d)** 15 months from the first presentation, the eye has converted to exAMD with new SRF. **e)** Risk prediction plot shows a steady increase in prediction score. At baseline, the prediction score was below both chosen operating points. Here, the majority of experts predicted imminent conversion in both the single and sequential tasks. At 7 months, the prediction score has risen above the liberal operating point, but still below the conservative operating point. Less than half of the experts predicted imminent conversion. At 10 months, 5 months before the eye converts, the system’s prediction rises above both operating points, correctly signalling that this eye is likely to convert to exAMD within the next 6 months.


a) Baseline
VA 76 letters
b) 10 months
VA 85 letters
c) 26 months
VA 82 letters
d) 42 months
VA 74 letters
e) 68 months
VA 63 letters
HRF
Drus.
PED
0.1 mm
0 mm
0.1 mm
0 mm
f)
System Prediction
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
4
1
4
1
6
3
0
3
6
9
12
15
18
21
24
27
30
33
36
39
42
45
48
51
54
57
60
63
66
69
72
Months since first presentation
Model prediction for conversion within 6 months
Threshold at conservative operating point
Threshold at liberal operating point
Single scan task: num. experts predicting conversion (n=6)
Sequential scan task: num. experts predicting conversion (n=6)

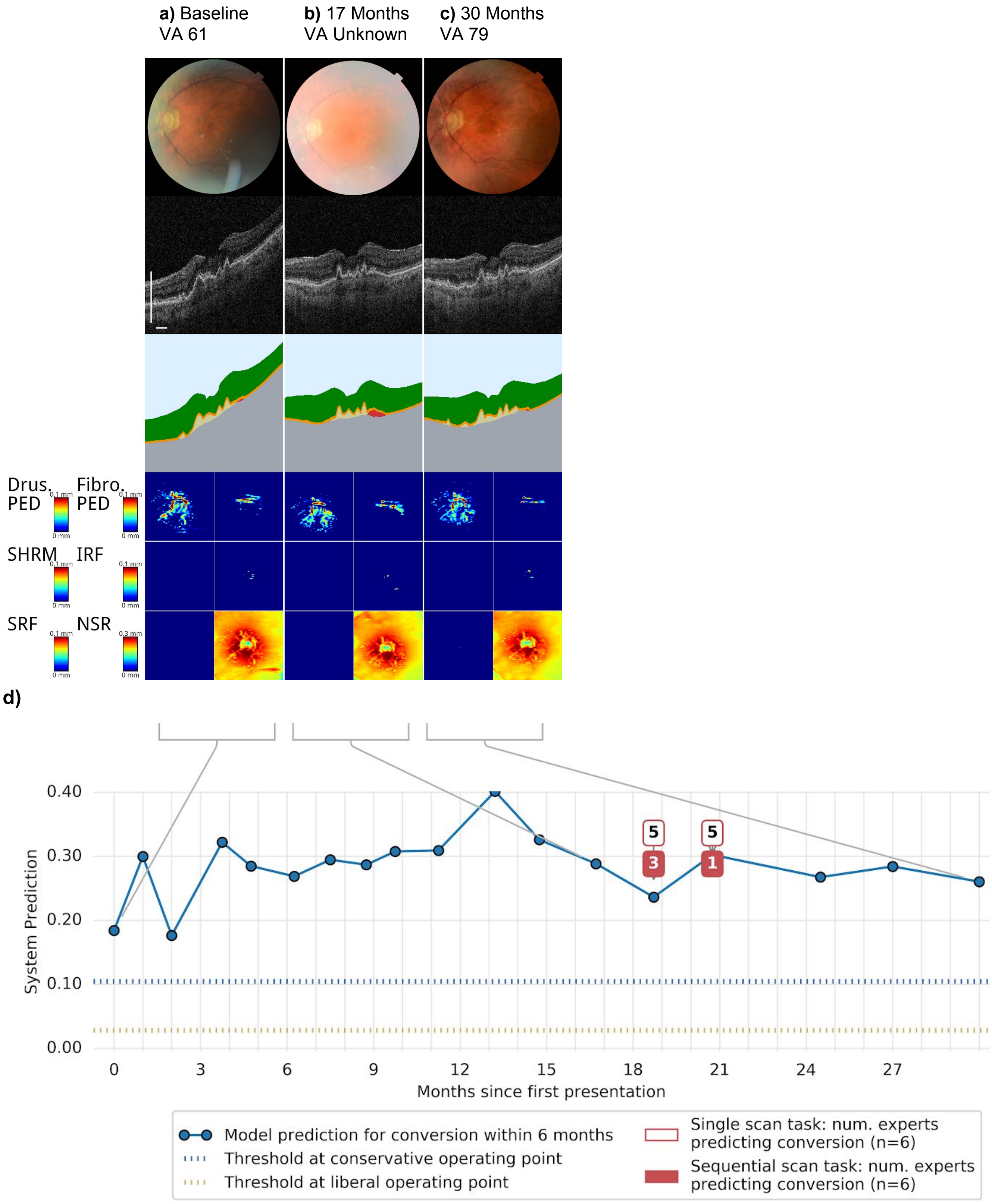
Supplementary Figure 3 | **A true negative success case of a 68 year old Female patient of 'other' ethnicity presenting with exAMD in the right eye**. The fellow eye, left eye, is shown in the figure. **a)** At baseline, large drusen are present, mostly focused around the central 3mm. HRF are present in the intraretinal tissues lying above the drusen. **b)** After 10 months, the drusen height and volume has increased slightly. **c)** 2 years from baseline, the height of the drusen regresses below the fovea, but the volume and number of drusen overall increases. **d)** Over 1 year later, a large drusenoid PED forms below the fovea, HRF increase in number, and the visual acuity drops from 82 to 74 letters. **e)** Another 2 years later, the drusenoid PED has grown larger, occupying a significant volume below the fovea. The visual acuity drops further to 63 letters. **f)** The risk-prediction plot correctly predicts that this eye will not convert at every visit despite the presence of large drusen and HRF. More than half of the experts predicted that the eye would convert imminently at all 3 scans presented (b, c, d) for the single scan task. When given all historical scans in the sequential task, fewer experts believed the eye would convert within 6 months.


a) Baseline
b) 35 months
c) 42 months
d) 44 months
e) 47 months
f) 66 months
Drus. PED
Fibro. PED
SHRM
IRF
SRF
NSR
0.1 mm
0 mm
0.3 mm
g)
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
System Prediction
4
3
6
2
4
3
First Injection
Conversion
0
3
6
9
12
15
18
21
24
27
30
33
36
39
42
45
48
51
54
57
60
63
66
69
Months since first presentation
Model prediction for conversion within 6 months
6 Months window prior to conversion scan
Threshold at conservative operating point
Threshold at liberal operating point
Single scan task: num. experts predicting conversion (n=6)
Sequential scan task: num. experts predicting conversion (n=6)

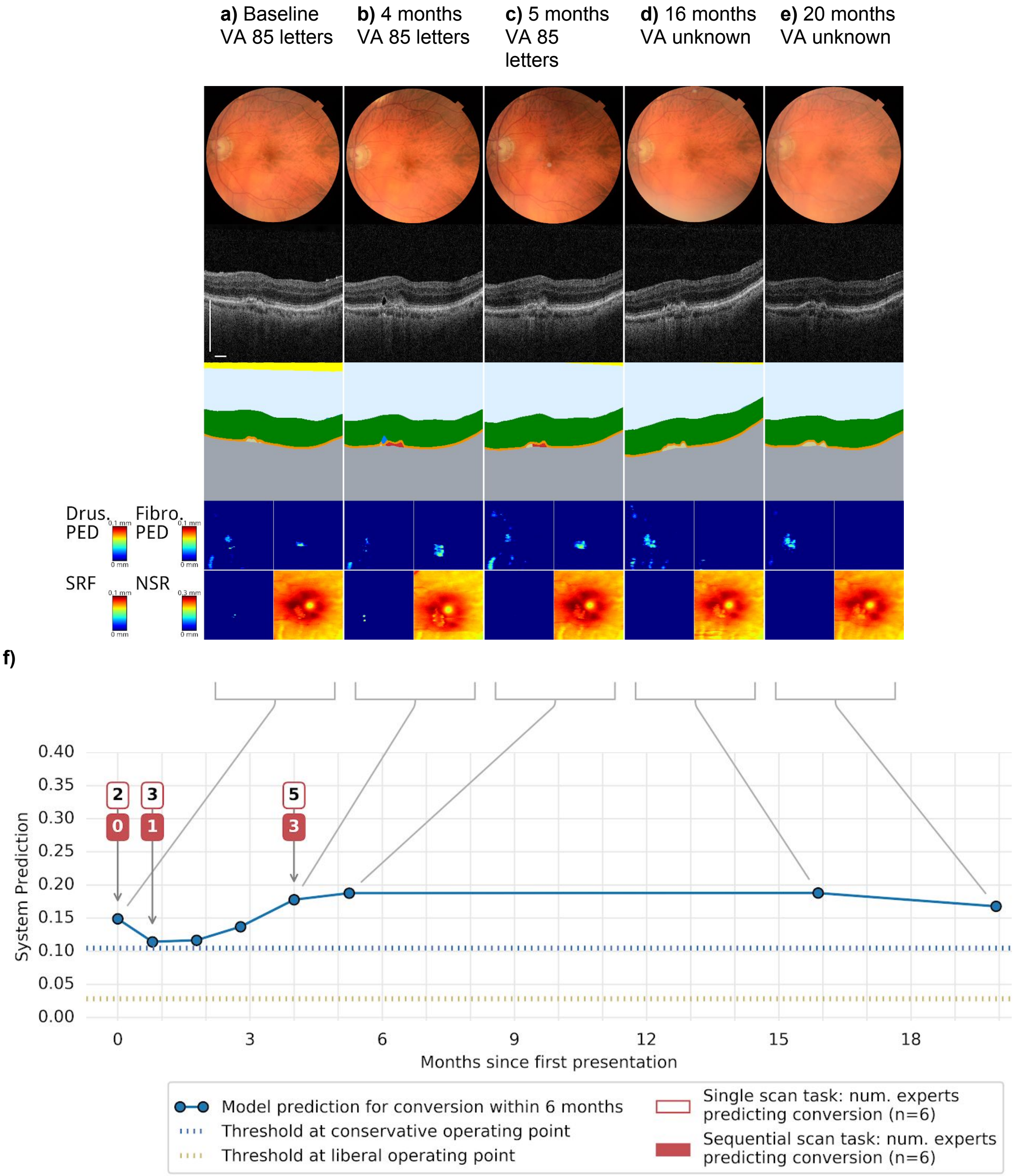
Supplementary Figure 4 | **A false negative failure case of an ambiguous case with questionable conversion to exAMD.** A 75 year old Indian female presented to the hospital eye service with approximately 6/18 vision in both eyes. The left eye was labelled as a converting fellow eye. However, there is a poor response to treatment with persistent fluid and PED. This is suspicious of simply a large drusenoid PED with overlying fluid, rather than true conversion to exAMD. In this case, the model predicts the eye will not convert with high confidence at all previous scans. **a)** The patient's first OCT scan of the left eye reveals confluent central drusen. A drusenoid PED is observed in the right eye (not shown). **b)** Nearly 3 years later, a large drusenoid PED is present. The segmentation model segments predominantly drusen material. At this point, the right eye converts with significant SRF. **c)** At 42 months the drusenoid PED has extended in diameter. **d)** Only 8 weeks later subretinal fluid appears above the PED for the first time. Here a consensus of conversion to exAMD was called. From this point on the segmentation model has some uncertainty trying to distinguish fibrovascular PED from drusenoid. **e)** The eye receives their first injection 4 weeks later, and after 3 injections at 47 months, the SRF appears persistent. **f)** Nearly 1.5 years later, the PED appears larger, with persistent SRF. There is a suboptimal response to treatment, and therefore may warrant further investigation. **g)** Risk prediction graph showing low risk of 6 month conversion up to conversion date, a false negative.


a) Baseline
VA unknown
b) 8 months
VA 84 letters
c) 10 months
VA 84 letters
Drus.
PED
Fibro.
PED
SHRM
NSR
0.1 mm
0 mm
0.3 mm
d)
0
0
0
0
0
0
Conversion
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
System Prediction
0
3
6
9
Months since first presentation
Model prediction for conversion within 6 months
6 Months window prior to conversion scan
Threshold at conservative operating point
Threshold at liberal operating point
Single scan task: num. experts predicting conversion (n=6)
Sequential scan task: num. experts predicting conversion (n=6)

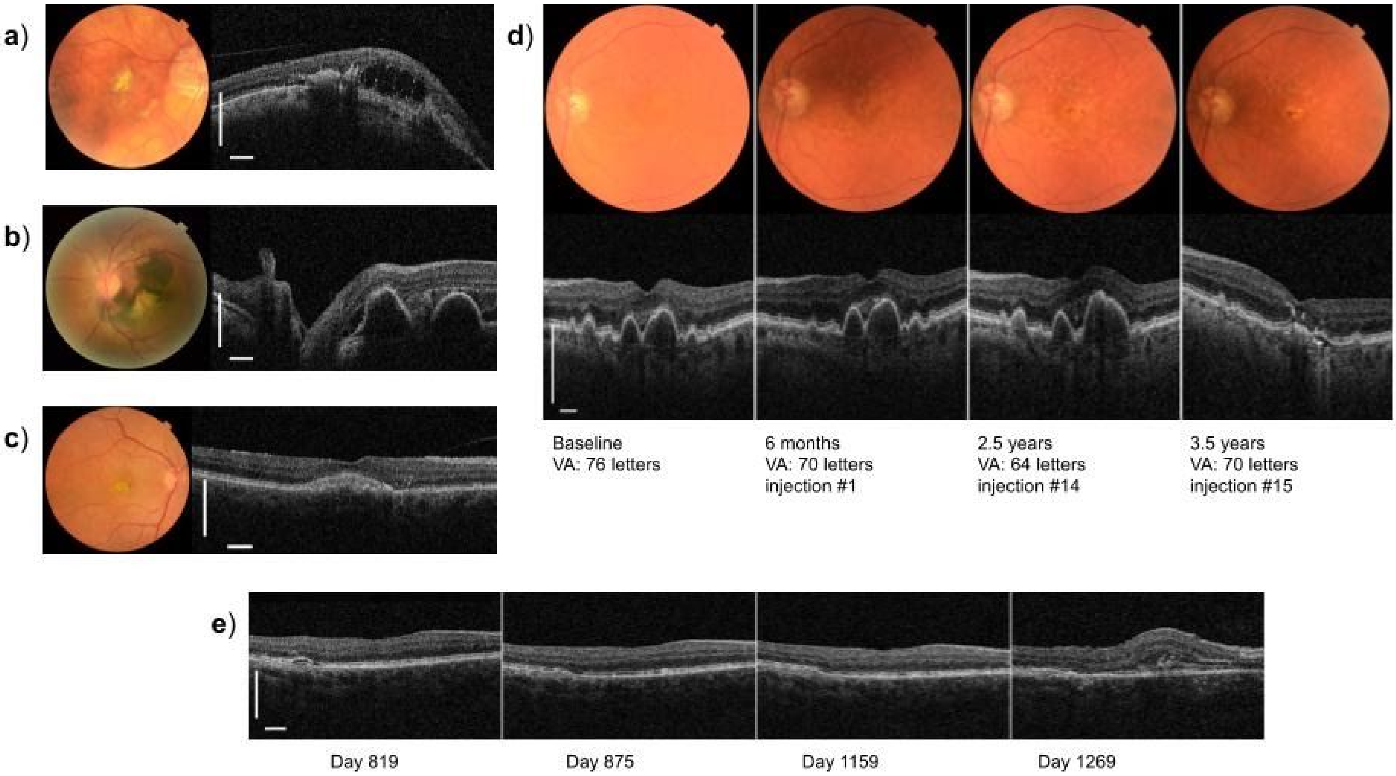
Supplementary Figure 5 | A false negative failure case of an 84 year old female white female was diagnosed with exAMD in the left eye, and who started treatment. Her (right) fellow eye was followed up and scanned at every follow-up visit. **a)** At presentation, the fellow eye has scattered drusen on OCT, appearing as white punctate accumulations on the fundus photo. **b)** After 8 months, there is little change in OCT appearance. **c)** 2 months later the eye converts to exAMD with the new appearance of SHRM with a fibrovascular PED. **d)** Risk prediction graph showing incorrect low risk of 6 month conversion up to conversion date. At all 3 scans presented, not a single expert predicted conversion within 6 months on both the single and sequential tasks.


a) Baseline
VA 61
b) 17 Months
VA Unknown
c) 30 Months
VA 79
Drus. PED
Fibro. PED
SHRM
IRF
SRF
NSR
0.1 mm
0 mm
0.3 mm
d)
0.40
0.30
0.20
0.10
0.00
System Prediction
0
3
6
9
12
15
18
21
24
27
Months since first presentation
5
3
5
1
Model prediction for conversion within 6 months
Threshold at conservative operating point
Threshold at liberal operating point
Single scan task: num. experts predicting conversion (n=6)
Sequential scan task: num. experts predicting conversion (n=6)

Supplementary Figure 6 | **A false positive failure case of suspicious subclinical choroidal neovascularisation.** This eye did not receive any treatment during the available follow-up of over 2 years. **a)** An irregular PED is observed on the first scan, suspicious of exAMD. As there is a lack of fluid, the eye is not labelled as exAMD at this point. The model segments this tissue as drusenoid PED with some elements of fibrovascular and believes there is a high risk that the eye has already converted. **b)** The segmentation model continues to segment a similar area of drusenoid and fibrovascular PED. **c)** The last scan available for this eye shows no significant change in the size of the PEDs and the enface maps look similar to the prior scans. **d)** Illustration of the risk-prediction. Here the system predicts that the eye will convert imminently: the system prediction for 6 month conversion is above the threshold for every visit. For the single-scan task, the experts were split between 'already converted', 'will convert imminently' and 'will not convert for >12 months'. When given historical scans in the sequential task, half of the experts believed the eye will convert. This is likely to be a case that would warrant further investigation such as OCT angiography imaging.


a) Baseline
VA 85 letters
b) 4 months
VA 85 letters
c) 5 months
VA 85 letters
d) 16 months
VA unknown
e) 20 months
VA unknown
Drus. PED
Fibro. PED
0.1 mm
0 mm
SRF
NSR
0.1 mm
0 mm
0.3 mm
0 mm
f)
2
0
3
1
5
3
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
System Prediction
0
3
6
9
12
15
18
Months since first presentation
Model prediction for conversion within 6 months
Threshold at conservative operating point
Threshold at liberal operating point
Single scan task: num. experts predicting conversion (n=6)
Sequential scan task: num. experts predicting conversion (n=6)

Supplementary Figure 7 | A false positive failure case demonstrating false positive predictions at every visit, with questionable low-grade exAMD. This patient is an 81 year old Black-Caribbean male who started treatment for exAMD in his right eye. His fellow eye (left eye) was labelled as a non-converting eye during the follow-up available. At every visit, the system predicted that the eye would convert imminently. **a)** His fellow eye, shown in the figure, presented with a small PED with fuzzy overlying RPE. **b**) After 4 months, a small pocket of SRF appeared above the PED. Here the segmentation model predicts that this is a fibrovascular PED. **c)** 1.5 months later, the SRF had resolved. **d)** 16 months and **e)** 20 months later, the PED appears to occupy a larger volume however there is no surrounding fluid. The segmentation model predicts this is a drusenoid PED. This particular case may represent a low-grade form of exAMD.



Supplementary Figure 8 | Difficult and ambiguous example cases. a) A 73-year-old white female presented with reduced vision. The fundus appearance is consistent with high myopia - choroidal vessels are visible and there is significant peripapillary atrophy. A circumscribed area of exudation is visible at the macula. The convexity of the macula is consistent with a dome-shaped macula, a condition associated with high myopia. Drusen was not seen in either eye. These findings likely represent myopic macular degeneration. **b)** A 59-year-old black female presented with a large peripapillary subretinal haemorrhage extending towards the fovea. OCT revealed multiple PEDs with overlying subretinal fluid adjacent to the optic nerve. The patient's fellow eye had no drusen changes. These features, and the ethnicity and age, suggest a variant of peripapillary choroidal neovascularisation or idiopathic polypoidal choroidal vasculopathy. **c)** This 87-year-old white female presented with an accumulation of yellow heterogeneous material located between the photoreceptor layer and the RPE, known as adult-onset vitelliform macular dystrophy. This condition is also age-related and is often mistaken for exAMD. The patient received 3 injections without any anatomical or visual response. **d)** A 75-year-old female who received treatment without clear evidence of exudative AMD. The patient presented with large intermediate drusen and a non-exudative detachment of the neurosensory retina - observed as SRF bridging between drusen. At 6 months the appearance was unchanged, but treatment was started as vision had reduced. The neurosensory detachment and drusenoid PED was still present 2 years after treatment. After 3 years of treatment and 14 injections, the drusenoid PED collapsed with little to no overlying SRF. **e)** Early appearance of SRF, temporal to the fovea, was observed in this 69-year-old male 819 days after starting treatment in the first eye. The SRF resolved 8 weeks later. After nearly 15 months, the eye converted to exAMD – the retina is thickened centrally, with SRF and SHRM adjacent to the previous area of fluid. The consensus grader label of conversion was at day 1269**.**

Supplementary Tables

Supplementary Table 1 | Difference in OCT and Segmentation model performance for selected examples of patient subgroups identified using automatic segmentation.

(a) OCT model performance for all subgroups were derived using automated segmentation of individual scans without exAMD in the test set.

Patient subgroup	Number of scans	Imminency scan prevalence (%)	Imminency AUC (95% CI)	Conservative operating point (90% specificity)		Liberal operating point (80% sensitivity)	
				Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)
All patients	5581	4.3	0.759 (0.731-0.787)	44.4	86.5	77.6	56.2
No drusen (no AMD)	425	0.0	N/A	N/A	98.1	N/A	95.1
Drusen volume 25th percentile	971	0.5	0.874 (0.866-0.954)	20.0	96.0	60.0	86.3
Drusen volume 25-50th percentile	1395	4.3	0.694 (0.620-0.768)	28.3	91.0	70.0	58.0
Drusen volume 50-75th percentile	1395	5.9	0.617 (0.550-0.684)	34.9	84.6	69.9	43.7
Drusen volume 75-100th percentile	1395	6.7	0.775 (0.722-0.828)	64.5	73.0	90.3	32.0
Geographic atrophy present	1573	4.0	0.702 (0.641-0.763)	39.7	79.3	87.3	38.4
Geographic atrophy absent	4008	4.4	0.787 (0.755-0.819)	46.1	89.4	74.2	63.2
HRF present	3867	5.3	0.743 (0.710-0.776)	51.5	81.8	83.5	43.9
HRF absent	1714	2.0	0.769 (0.721-0.817)	2.9	96.7	42.9	83.1
Fibrovascular PED present	2326	6.6	0.707 (0.661-0.753)	58.4	72.2	87.0	28.3
Fibrovascular PED absent	3255	2.7	0.781 (0.743-0.819)	19.5	96.3	60.9	75.4

(b) Segmentation model performance for all subgroups were derived using automated segmentation of individual scans without exAMD in the test set.

				Conservative operating point (90% specificity)		**Liberal operating point (80% sensitivity)**	
Patient subgroup	**Number of scans**	**Imminency scan prevalence (%)**	**Imminency AUC (95% CI)**	**Sensitivit y (%)**	**Specificity (%)**	**Sensitivity (%)**	**Specificity (%)**
All patients	5581	4.3	0.731 (0.704-0.758)	32.0	90.0	80.1	53.1
No drusen (no AMD)	425	0.0	N/A	N/A	100.0	N/A	94.6
Drusen volume 25th percentile	971	0.5	0.914 (0.851-0.977)	20.0	98.3	80.0	87.9
Drusen volume 25-50th percentile	1395	4.3	0.666 (0.603-0.729)	10.0	94.6	63.3	60.8
Drusen volume 50-75th percentile	1395	5.9	0.566 (0.501-0.631)	26.5	86.1	75.9	31.9
Drusen volume 75-100th percentile	1395	6.7	0.733 (0.684-0.782)	51.6	79.7	94.6	27.3
Geographic atrophy present	1573	4.0	0.680 (0.620-0.740)	28.6	87.3	90.5	32.3
Geographic atrophy absent	4008	4.4	0.760 (0.728-0.792)	33.1	91.1	76.4	61.4
HRF present	3867	5.3	0.706 (0.672-0.740)	37.4	86.7	87.9	39.7
HRF absent	1714	2.0	0.772 (0.726-0.818)	0.0	97.2	34.3	82.5
Fibrovascular PED present	2326	6.6	0.647 (0.607-0.687)	44.8	77.0	92.2	21.2
Fibrovascular PED absent	3255	2.7	0.771 (0.731-0.811)	9.2	98.9	58.6	75.0

Supplementary Table 5 | Test set imminency performance with AUC weighting. Weighting is used to average the number of scans per patient at liberal and conservative operating points. N=5581.

Operating point	**Sensitivity (95% CI)**	**Specificity (95% CI)**
	Weighted Imminency	
Liberal	0.800 (0.774-0.826)	0.525 (0.422-0.628)
Conservative	0.393 (0.292-0.494)	0.900 (0.889-0.901)
	Unweighted Imminency	
Liberal	0.800 (0.750-0.850)	0.551 (0.495-0.607)
Conservative	0.344 (0.2820.406)	0.900 (0.889-0.901)

Supplementary Table 3 | Model and ensemble performance on cross-validation folds. Performance is shown for the task of prediction of exAMD within 6 months (imminency). Each model type's outputs were averaged over three replicas with different random initialization and small amounts of TTA applied (see Methods for more details). The mean (m) and standard deviation (std) among the replicas are given in parentheses. The best performing model type was then used for test time evaluation which was determined to be the Ensemble system from ensembling over OCT and segmentation models. Cross-validation folds were generated at the patient level to ensure approximately equal patient representation. Folds were generated before any exclusion criterias resulting in slightly different counts for each fold.

Fold	Non-excluded patients	Imminency AUC		
		OCT model	Segmentation model	Ensemble system
1	601	0.81 (m=0.80, std=0.002)	0.81 (m=0.80, std=0.007)	**0.82**
2	595	0.75 (m=0.74, std=0.011)	**0.78 (m=0.77, std=0.008)**	**0.78**
3	610	**0.80 (m=0.79, std=0.011)**	0.79 (m=0.78, std=0.007)	**0.81**
4	603	**0.78 (m=0.77, std=0.008)**	0.76 (m=0.75, std=0.001)	**0.78**

Supplementary Table 7 | False positive alerts. a) Number of eyes with at least one false positive (FP) alert with varying operating points, stratified by converting and non-converting fellow eyes, and the follow-up available for the eye. Highlighted rows correspond to the liberal (80% sensitivity) and conservative (90% specificity) thresholds.
b) Number of eyes with at least one false positive (FP) alert at the liberal (80% sensitivity) and conservative (90% specificity) thresholds, those with 24 months follow-up from the FP scan, and the number (%) which were non-converting throughout the study, those converting with the FP 6-24 months prior to conversion, and those with the FP >24 months prior to conversion.

a)

			Unique converting eyes with at least one FP (n=103)			**Unique non-converting eyes with at least one FP (n=283)**		
Sensitivity	**Specificity**	**Total unique eyes with at least one FP (n=386)**	**6-12 months prior to conversion**	**6-24 months prior to conversion**	**>6 months prior to conversion**	**≥6 months follow-up**	**≥12 months follow-up**	**≥24 months follow-up**
9.54%	98.45%	20	1	4	4	16	13	8
20.33%	96.42%	38	9	10	10	28	25	16
29.88%	93.02%	66	13	16	17	49	38	23
39.83%	87.06%	103	18	23	26	77	69	43
49.79%	78.41%	157	31	40	43	114	98	70
60.17%	72.25%	170	37	45	48	122	114	79
70.12%	64.63%	201	42	50	56	145	133	95
80.08%	54.55%	237	49	58	62	175	163	115
90.04%	40.37%	264	54	66	71	193	183	135
100.00%	16.50%	323	60	70	76	247	235	171
Sensitivity	**Specificity**							
99.59%	**16.50%**	323	60	70	76	247	235	171
98.76%	**21.12%**	310	60	70	76	234	220	165
96.68%	**30.11%**	292	59	70	75	217	202	148
90.04%	**39.89%**	265	54	66	71	194	184	135
83.40%	**50.11%**	246	51	59	63	183	173	124
73.86%	**60.02%**	218	45	53	58	160	147	102
62.66%	**69.14%**	180	38	45	50	130	118	83
46.89%	**79.93%**	147	29	36	40	107	94	68
34.02%	**89.93%**	88	16	19	21	67	56	35
0.00%	**100.00%**	0	0	0	0	0	0	0

b)

	Total unique eyes with at least one FP (n=386)	**Total unique eyes with at least one FP and 24 months follow-up**	**Non-converting**	**6-24 months prior to conversion**	**>24 months prior to conversion**
Liberal Operating Point	236	174	116 (66.7%)	57 (32.8%)	27 (15.5%)
Conservative Operating Point	89	54	35 (64.8%)	19 (35.2%)	6 (11.1%)

Supplementary Table 6 | System performance with varying lead times. System performance is shown for prediction of conversion within 3 month, 6 month, 12 month, and 24 month time windows at a scan-level. Sensitivity and specificity thresholds were varied from 10% (or lowest possible) to 100%.

a) System performance predicting future conversion to exAMD within 3 months

Sensitivity	Specificity	TP	FP	TN	FN	PPV	NPV	Sensitivity	Specificity	TP	FP	TN	FN	PPV	NPV
10.09%	98.23%	10	103	5723	99	0.088	0.983	99.08%	**17.95%**	108	4779	1047	1	0.022	0.999
20.18%	96.72%	21	191	5635	88	0.099	0.985	99.08%	**17.95%**	108	4779	1047	1	0.022	0.999
30.28%	93.65%	32	370	5456	77	0.080	0.986	97.25%	**33.02%**	106	3901	1925	3	0.026	0.998
40.37%	89.58%	43	607	5219	66	0.066	0.988	95.41%	**39.84%**	104	3504	2322	5	0.029	0.998
49.54%	81.82%	53	1059	4767	56	0.048	0.988	88.99%	**50.02%**	97	2911	2915	12	0.032	0.996
59.63%	74.48%	64	1487	4339	45	0.041	0.990	82.57%	**59.89%**	90	2336	3490	19	0.037	0.995
69.72%	68.47%	75	1837	3989	34	0.039	0.992	66.06%	**70.01%**	72	1746	4080	37	0.040	0.991
79.82%	62.44%	86	2188	3638	23	0.038	0.994	51.38%	**79.83%**	56	1174	4652	53	0.046	0.989
89.91%	50.02%	97	2912	2914	12	0.032	0.996	38.53%	**90.20%**	42	570	5256	67	0.069	0.987
100.00%	17.95%	108	4780	1046	1	0.022	0.999	0.00%	**100.00%**	0	0	5826	109	NaN	0.982

b) System performance predicting future conversion to exAMD within 6 months

Sensitivity	Specificity	TP	FP	TN	FN	PPV	NPV	Sensitivity	Specificity	TP	FP	TN	FN	PPV	NPV
9.54%	98.45%	22	83	5257	219	0.210	0.960	99.59%	**16.50%**	240	4458	882	1	0.051	0.999
20.33%	96.42%	48	191	5149	193	0.201	0.964	98.76%	**21.12%**	238	4211	1129	3	0.053	0.997
29.88%	93.02%	71	373	4967	170	0.160	0.967	96.68%	**30.11%**	233	3731	1609	8	0.059	0.995
39.83%	87.06%	95	691	4649	146	0.121	0.970	90.04%	**39.89%**	217	3209	2131	24	0.063	0.989
49.79%	78.41%	119	1153	4187	122	0.094	0.972	83.40%	**50.11%**	201	2663	2677	40	0.070	0.985
60.17%	72.25%	144	1482	3858	97	0.089	0.975	73.86%	**60.02%**	178	2134	3206	63	0.077	0.981
70.12%	64.63%	168	1889	3451	73	0.082	0.979	62.66%	**69.14%**	151	1647	3693	90	0.084	0.976
80.08%	54.55%	192	2427	2913	49	0.073	0.983	46.89%	**79.93%**	113	1071	4269	128	0.095	0.971
90.04%	40.37%	216	3184	2156	25	0.064	0.989	34.02%	**89.93%**	82	537	4803	159	0.132	0.968
100.00%	16.50%	240	4459	881	1	0.051	0.999	0.00%	**100.00%**	0	0	5340	241	NaN	0.957

c) System performance predicting future conversion to exAMD within 12 months

Sensitivity	Specificity	TP	FP	TN	FN	PPV	NPV	Sensitivity	Specificity	TP	FP	TN	FN	PPV	NPV
10.32%	98.00%	48	89	4359	427	0.350	0.911	99.79%	**14.86%**	474	3786	662	1	0.111	0.998
20.00%	95.17%	94	215	4233	381	0.304	0.917	98.95%	**19.45%**	470	3582	866	5	0.116	0.994
29.89%	89.30%	141	476	3972	334	0.229	0.922	96.42%	**29.99%**	458	3113	1335	17	0.128	0.987
40.00%	83.12%	189	751	3697	286	0.201	0.928	88.21%	**40.42%**	419	2649	1799	56	0.137	0.970
49.89%	75.97%	236	1069	3379	239	0.181	0.934	81.05%	**49.82%**	385	2231	2217	90	0.147	0.961
60.00%	69.24%	284	1368	3080	191	0.172	0.942	72.63%	**60.07%**	345	1775	2673	130	0.163	0.954
69.89%	62.19%	331	1682	2766	144	0.164	0.951	57.89%	**69.96%**	275	1335	3113	200	0.171	0.940
80.00%	53.73%	379	2058	2390	96	0.156	0.961	44.00%	**79.92%**	209	892	3556	266	0.190	0.930
90.11%	37.63%	427	2774	1674	48	0.133	0.972	28.63%	**89.97%**	136	445	4003	339	0.234	0.922
100.00%	14.86%	474	3787	661	1	0.111	0.998	0.00%	**100.00%**	0	0	4448	475	NaN	0.904

d) System performance predicting future conversion to exAMD within 24 months

Sensitivity	Specificity	TP	FP	TN	FN	PPV	NPV	Sensitivity	Specificity	TP	FP	TN	FN	PPV	NPV
10.20%	97.07%	81	79	2613	723	0.506	0.783	99.75%	**9.32%**	802	2440	252	2	0.247	0.992
20.02%	93.91%	160	164	2528	644	0.494	0.797	99.01%	**18.50%**	796	2193	499	8	0.266	0.984
29.85%	90.01%	239	269	2423	565	0.470	0.811	97.89%	**30.35%**	787	1874	818	17	0.296	0.980
39.93%	86.59%	320	361	2331	484	0.470	0.828	88.56%	**40.12%**	712	1611	1081	92	0.307	0.922
50.00%	79.75%	401	545	2147	403	0.424	0.842	81.09%	**50.04%**	652	1344	1348	152	0.327	0.899
59.95%	71.10%	481	778	1914	323	0.382	0.856	72.76%	**60.03%**	585	1075	1617	219	0.352	0.881
70.02%	62.52%	562	1009	1683	242	0.358	0.874	61.19%	**70.02%**	492	806	1886	312	0.379	0.858
79.98%	52.34%	642	1283	1409	162	0.334	0.897	49.50%	**80.05%**	398	536	2156	406	0.426	0.842
90.05%	38.00%	723	1669	1023	81	0.302	0.927	29.73%	**90.01%**	239	268	2424	565	0.471	0.811
100.00%	4.83%	803	2562	130	1	0.239	0.992	0.00%	**100.00%**	0	0	2692	804	NaN	0.770

Supplementary Table 8 | Datasets used. An overview of all datasets used for training, validation and testing of the different networks.

Dataset		Number of scans	Input	Labels	Label source
#1	Training set for segmentation	846	OCT scans	Sparse segm. maps (3-5 slices per scan)	Manually segmented by trained ophthalmologists, reviewed and edited by a senior ophthalmologist. Used for the original model described by De Fauw et al. 2018.
#2	Validation set for segmentation	181	OCT scans	Sparse segm. maps (3-5 slices per scan)	Manually segmented by trained ophthalmologists, reviewed and edited by a senior ophthalmologist. Used for the original model described by De Fauw et al. 2018.
#3	Overridden training/validation set for segmentation	421	OCT scans	Sparse segm. maps (3-5 slices per scan)	Manually segmented images from sets #1 and #2 with HRF additionally segmented by trained optometrists and ophthalmologists, reviewed and edited by a senior optometrist.
#4	Additional training/validation set for segmentation	86	OCT scans	Sparse segm. maps (3-5 slices per scan)	Additional manually segmented images including extra examples of dry AMD and HRF by trained optometrists and ophthalmologists, reviewed and edited by a senior optometrist.
#5	Training set for classification loss in predictive model	35,575	OCT scans	Diagnoses and referral decision	Automated notes search + trained ophthalmologist and optometrist review of the OCT scans.
#6	Training set for predictive model	67,802	OCT scans, segmentation maps	Date of conversion	Graded by two graders. Disagreement in conversion labels arbitrated by a senior grader.
#7	Validation set for predictive model	22,583	OCT scans, segmentation maps	Date of conversion	Graded by two graders. Disagreement in conversion labels arbitrated by a senior grader.
#8	Hold-out test set	5,581	OCT scans	Date of conversion	Graded by two graders. Disagreement in conversion labels arbitrated by a senior grader.
#9	Benchmark study: Single scan task	1,053	Subset of scans in #8	Individual predictions from 6 experts	6 experts (3 retinal specialists and 3 optometrists) grading on OCT scan only
#10	Benchmark study: Sequential task	20,706	Same scans as #9 (plus all historical scans for both eyes)	Individual predictions from 6 experts	6 experts (3 retinal specialists and 3 optometrists) grading on OCT scan, fundus image and clinical notes

Supplementary Table 11 | Subgroup derivation using OCT segmentation maps.

Non-converted scan feature	**Definition**
No drusen	Absence of drusen
Any drusen	Drusenoid PED of any size
Drusen volume 0-25th percentile	Drusen volume between 1 and 4124 voxels (0 and 5.9e^{-3} mm^{3})
Drusen volume 25-50th percentile	Drusen volume between 4125 and 22760 voxels (5.9e^{-3} and 3.3e^{-2} mm^{3})
Drusen volume 50-75th percentile	Drusen volume between 22761 and 72070 voxels (3.3e^{-2} and 0.1 mm^{3})
Drusen volume 75-100th percentile	Drusen volume between 72071 and 7.5e^{5} voxels (0.1 and 1.1 mm^{3})
Geographic atrophy (GA)	At least one atrophic area with diameter ≥ 250 μm
Hyperreflective foci (HRF)	Present if at least one instance present with ≥4 voxels in volume
Conversion scan feature	
Fibrovascular pigment epithelial detachment (PED)	Present if ≥ 5 voxels in volume overall
Serous PED	Present if ≥ 5 voxels in volume overall
Subretinal hyperreflective material (SHRM)	Present if at least one area ≥ 5 voxels in volume
Intraretinal fluid (IRF)	Present if at least one cyst with ≥ 5 voxels in volume
Subretinal fluid (SRF)	Present if at least one area ≥ 5 voxels in volume

Supplementary Table 12 | System performance across different subgroups

a) System performance across subgroups derived using segmentation

							Conservative operating point (90% specificity)		Liberal operating point (80% sensitivity)	
Tissue presence	**HRF presence**	**Number of scans**	**Number of eyes**	**Converting eyes (%)**	**Imminency prevalence (%)**	**Imminency AUC (95% CI)**	**Sensitivity (%)**	**Specificity (%)**	**Sensitivity (%)**	**Specificity (%)**
All patients	All	5581	386	103 (26.7)	4.3	0.745 (±0.027)	33.6	90.0	79.7	55.1
	Present	3867	365	97 (26.6)	5.3	0.725 (±0.033)	38.8	86.7	87.9	41.7
	Absent	1714	227	48 (21.1)	2.0	0.779 (±0.042)	2.9	97.3	31.4	84.6
Any drusen	All	5156	376	103 (27.4)	4.7	0.725 (±0.031)	33.6	89.2	79.7	51.7
	Present	3761	356	97 (27.2)	5.5	0.717 (±0.035)	38.8	86.3	87.9	40.1
	Absent	1395	216	48 (22.2)	2.5	0.739 (±0.05)	2.9	96.7	31.4	82.1
No drusen (no AMD)	All	425	41	0 (0.0)	0.0	n/a	0.0	100.0	0.0	95.1
	Present	106	31	0 (0.0)	0.0	n/a	0.0	100.0	0.0	94.3
	Absent	319	38	0 (0.0)	0.0	n/a	0.0	100.0	0.0	95.3
Drusen volume 25th percentile	All	1396	135	12 (8.9)	0.4	0.931 (±0.042)	0.0	98.5	60.0	90.4
	Present	551	114	10 (8.8)	0.9	0.866 (±0.080)	0.0	96.3	60.0	82.1
	Absent	845	98	7 (7.1)	0.0	n/a	0.0	99.9	0.0	95.9
Drusen volume 25-50th percentile	All	1395	172	48 (27.9)	4.3	0.681 (±0.065)	15.0	94.8	68.3	61.9
	Present	962	158	44 (27.8)	4.7	0.679 (±0.072)	17.8	93.1	82.2	50.6
	Absent	433	87	21 (24.1)	3.5	0.706 (±0.006)	6.7	98.3	26.7	86.6
Drusen volume 50-75th percentile	All	1395	186	60 (32.3)	5.9	0.584 (±0.067)	24.1	87.3	72.3	36.3
	Present	1130	172	53 (30.8)	5.7	0.638 (±0.077)	31.3	86.3	84.4	30.6
	Absent	265	78	23 (29.5)	7.2	0.511 (±0.009)	0.0	91.9	31.6	61.0
Drusen volume 75-100th percentile	All	1395	144	46 (31.9)	6.7	0.759 (±0.049)	55.9	78.8	94.6	29.6
	Present	1224	139	46 (33.1)	7.5	0.746 (±0.051)	56.5	77.1	94.6	25.4
	Absent	171	38	12 (31.6)	0.6	0.812 (±0.057)	0.0	90.0	100.0	57.6
Geographic atrophy present	All	1573	206	52 (25.2)	4.0	0.692 (±0.059)	31.7	85.7	88.9	34.9
	Present	1408	188	50 (26.6)	4.3	0.685 (±0.059)	32.8	84.3	90.2	31.6
	Absent	165	70	8 (11.4)	1.2	0.552 (±0.035)	0.0	97.5	50.0	62.0
Geographic atrophy absent	All	4008	327	87 (26.6)	4.4	0.774 (±0.032)	34.3	91.7	76.4	63.1
	Present	2459	305	81 (26.6)	5.9	0.748 (±0.038)	41.4	88.1	86.9	47.5
	Absent	1549	204	46 (22.5)	2.1	0.809 (±0.039)	3.0	97.3	30.3	87.0
Fibrovascular PED present	All	2326	280	90 (32.1)	6.6	0.675 (±0.043)	48.1	77.3	90.9	22.4
	Present	2036	259	85 (32.8)	7.2	0.68 (±0.047)	50.0	76.1	92.5	19.9
	Absent	290	96	28 (29.2)	2.8	0.481 (±0.088)	12.5	85.5	62.5	39.0
Fibrovascular PED absent	All	3255	298	64 (21.5)	2.7	0.784 (±0.038)	8.0	98.7	59.8	77.6
	Present	1831	269	56 (20.8)	3.3	0.757 (±0.054)	11.7	97.9	76.7	64.8
	Absent	1424	188	32 (17.0)	1.9	0.852 (±0.039)	0.0	99.7	22.2	93.8

b) System performance across subgroups based on tissue present on conversion scan (tissue presence threshold is 5 voxels in tissue segmentation volume)

Tissue presence on conversion scan	Number of converting eyes	Number of scans in history of eye	Imminency prevalence (%)	Imminency AUC (95% CI)	Conservative operating point (90% specificity)		Liberal operating point (80% sensitivity)	
					Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)
All	103	1119	21.5	0.626 (±0.04)	33.6	85.2	79.7	32.2
IRF	62	709	20.5	0.674 (±0.054)	44.8	80.7	89.7	26.4
SRF	71	813	19.3	0.635 (±0.051)	36.3	84.6	77.1	35.4
SHRM	69	715	19.9	0.626 (±0.057)	33.8	84.3	85.2	26.5
Fibrovascular PED	99	1045	22.2	0.634 (±0.041)	34.9	84.9	79.7	33.6
Serous PED	15	142	23.9	0.570 (±0.018)	32.4	63.9	64.7	37

c) System performance across subgroups based on outputs from the classification model described in De Fauw et al. (2018). CNV: choroidal neovascularisation, GA: geographic atrophy, VMT: vitreomacular traction, ERM: epiretinal membrane. Drusen and GA were classified as present if the classification network risk prediction was >0.9. CNV, VMT, and ERM were classified as present if the risk prediction was >0.8.

Referral and Disease outputs	Number of scans	Number of eyes	Converting eyes (%)	Imminency prevalence (%)	Imminency AUC (95% CI)	Conservative operating point (90% specificity)		Liberal operating point (80% sensitivity)	
						Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)
All patients	5581	386	103 (26.7)	4.3	0.745 (±0.027)	33.6	90	79.7	55.1
CNV present (P>0.8)	157	37	15 (40.5)	10.2	0.635 (±0.033)	75	41.8	100	0
CNV absent (P<0.8)	5424	381	98 (25.7)	4.1	0.744 (±0.03)	30.7	91.3	78.2	56.6
Drusen present (P>0.9)	3944	311	97 (31.2)	5.2	0.700 (±0.036)	29.4	90.7	79.9	47.7
Drusen absent (P<0.9)	1637	180	31 (17.2)	2.3	0.865 (±0.036)	56.8	88.5	78.4	72.6
GA present (P>0.9)	1528	168	44 (26.2)	3.7	0.725 (±0.057)	25	91.7	92.9	38.6
GA present, drusen present	1453	164	43 (26.2)	3.9	0.732 (±0.059)	25	91.6	92.9	39.8
GA absent (P<0.9)	4053	345	95 (27.5)	4.6	0.763 (±0.031)	36.2	89.4	75.7	61.5
GA absent, drusen absent	1562	175	31 (17.7)	2.4	0.872 (±0.032)	56.8	88.3	78.4	75.3
VMT (P>0.8)	84	19	2 (10.5)	1.2	1.00 (±0.000)	100	94	100	88
ERM (P>0.8)	551	59	12 (20.3)	3.1	0.772 (±0.002)	5.9	98.9	70.6	71.7

Supplementary Table 13 | System performance across demographics. Sensitivity and specificity at both conservative (90% specificity) and liberal (80% sensitivity) operating points. Performance is given at a scan-level.

Demographic		Number of scans	Number of eyes	Converting eyes (%)	Imminency prevalence (%)	Imminency AUC (95% CI)	Conservative operating point (90% specificity)		Liberal operating point (80% sensitivity)	
							Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)
All		**5581**	**386**	**103 (26.7)**	**4.3**	0.745 (0.718-0.772)	**33.6**	**90.0**	**79.7**	**55.1**
Sex	Female	3397	240	68 (28.3)	4.8	0.733 (0.696-0.770)	33.5	89.1	81.1	51.7
	Male	2184	146	35 (24.0)	3.5	0.764 (0.717-0.811)	33.8	91.4	76.6	60.4
Age	50-59	103	8	0 (0.0)	0.0	n/a	n/a	100.0	0.0	97.1
	60-69	734	59	10 (16.9)	1.2	0.751 (0.700-0.802)	0.0	96.0	77.8	71.3
	70-79	1769	156	34 (21.8)	3.1	0.722 (0.657-0.787)	31.5	93.4	61.1	63.3
	80-89	2413	211	66 (31.3)	5.6	0.709 (0.665-0.753)	37.8	84.9	82.2	45.0
Ethnicity	White	3010	206	61 (29.6)	4.7	0.732 (0.694-0.770)	36.6	88.9	80.3	50.8
	Black	139	10	3 (30.0)	4.3	0.954 (0.917-0.991)	16.7	95.5	100.0	85.0
	Asian	644	42	10 (23.8)	3.9	0.778 (0.694-0.862)	28.0	92.7	76.0	62.7
	Other	1141	89	24 (27.0)	4.4	0.729 (0.672-0.786)	30.0	89.1	72.0	59.1
	Unknown	647	39	5 (12.8)	2.8	0.780 (0.700-0.788)	33.3	92.7	94.4	54.2

Supplementary Table 14 | Overview of inputs used in this study. All OCT sizes are given in A-scan, B-scan, C-scan direction

Dataset	Image size [voxels]	Real-world voxel size [μm]	Real-World image size [mm]	Comments
Raw OCT scans	885 · 512 · 128	2.6 · 11.7 · 47.2	2.3 · 6.0 · 6.0	Images acquired on Topcon 3D OCT-2000 device
Segmentation network input	448 · 512 · 128	5.2 · 11.7 · 47.2		Raw OCT scans resampled in A-scan direction to 5.2μm voxel size, and zero-padded to the next multiple of 64 (added 6 pixels).
Clinical referral segmentation input	300 · 350 · 43	7.8 · 17.6 · 141.7		Segmentation map resampled to 7.8μm · 17.6μm · 141.7μm voxel size such that the full classification network fits into GPU memory
exAMD prediction OCT/segmentation input	450 · 450 · 41	5.7 · 17.6 · 141.7		Image and voxel size chosen as a tradeoff of GPU memory constraints and validation performance.

Supplementary Table 17 | Total OCT scans in the dataset stratified by time-to-conversion of the fellow eye.

Fellow eye conversion	Number of scans				
	Training	Validation	Test	Benchmark Study	Total
Converts within 0-6 months	846	281	240	141	**1,367**
Converts within 6-12 months	500	197	170	31	**867**
Converts after 12 months	1,901	589	662	63	**3,152**
Does not convert within study period with >6 months follow-up	14,980	5,084	4,462	808	**24,526**
Does not convert within study period with >12 months follow-up	12,580	4,265	3,784	613	**20,629**

Supplementary Table 16 | exAMD prediction network architecture details. Shapes are all given as [x, y, z]. Depending on the input modality, the channel dimension of the first two levels differ as denoted in parentheses (channels with OCT input, channels with segmentation input). The first 1x1x1 convolution in level 3 standardizes the shape. Block convolutions describe the list of convolutions and their kernels used in each block[1]. See Extended Data Figure 10 for a description of the block architecture.

Level	Type	1x1x1 and Max pool Kernel shape/ stride	Output shape	Block convolution layer list	Channels per block convolution
Input			450x450x41x(1,17)		
1	Block			[3x3x1, 3x3x1]	8
	Max pool	[2x2x1]/[2x2x1]	225x225x41x(17,33)		
2	Block			[3x3x1, 3x3x1, 1x1x3] x 2	16
	Max pool	[2x2x1]/[2x2x1]	112x112x41x(113,129)		
3	Conv	[1x1x1]/[1x1x1]			128
	Block			[3x3x1, 3x3x1, 1x1x3] x 2	16
	Max pool	[2x2x2]/[2x2x2]	56x56x20x224		
4	Conv	[1x1x1]/[1x1x1]			128
	Block			[3x3x1, 3x3x1, 1x1x3] x 2	32
	Block			[3x3x1, 3x3x1, 1x1x3] x 2	32
	Max pool	[2x2x2]/[2x2x2]	28x28x10x512		
5	Conv	[1x1x1]/[1x1x1]			128
	Block			[3x3x1, 3x3x1, 1x1x3] x 2	32
	Block			[3x3x1, 3x3x1, 1x3x3] x 2	32
	Max pool	[2x2x2]/[2x2x2]	14x14x5x512		
6	Conv	[1x1x1]/[1x1x1]			256
	Block			[3x3x1, 3x3x1, 1x1x3] x 2	32
	Block		14x14x5x640	[3x3x1, 3x3x1, 1x1x3] x 2	32
Output	Conv	[1x1x1]/[1x1x1]	14x14x5x128		128
	Global Avg pool	[14x14x5]/[14x14x5]	1		

[1] For example, the block in level 2 uses 6 convolution layers each with 16 channels. As described in Extended Data Figure 7, the convolutions are identical except for the kernels used which for this block are ordered as: [3x3x1]→[3x3x1]→[1x1x3]→[3x3x1]→[3x3x1]→[1x1x3]. Thus each block is fully described with the stack of kernels used and the channels per convolution.

Supplementary Table 4 | Gradient Boosted Machine baseline performances. To investigate if the model was overfitting to demographic information we trained a baseline Gradient Boosting Machine model using only the demographic metadata available in our dataset (sex, ethnicity, visual acuity, and age), data that was not used in training for the main model presented in this paper. A further experiment used tissue volumes derived from the segmentation model as features in addition to demographic metadata and visual acuity. Both methods overfit strongly and perform worse than the DLS on the cross validation folds, suggesting that demographic information alone is insufficient to predict conversion to exudative AMD. See **Supplementary Table 3** for patient and scan information on each fold.

Fold	**Demographic metadata and visual acuity only**		**Tissue volumes, demographic metadata and visual acuity**	
	Train imminency AUC	**Valid imminency AUC**	**Train imminency AUC**	**Valid imminency AUC**
1	0.73	0.58	0.88	0.74
2	0.75	0.52	0.88	0.73
3	0.72	0.51	0.88	0.75
4	0.73	0.59	0.89	0.70

Supplementary Table 9 | Model F1 score compared with human experts. F1 score at equal error point compared with human experts on the human benchmark task. The model performs better than all experts with statistical significance to five out of six experts (in bold). N=1053. P-values obtained with two-sided permutation tests.

	F1 score	**Difference with equal error point**
Equal error point	0.38	
Retina Specialist 1	0.29	-0.09 (p<0.01)
Retina Specialist 2	0.23	-0.15 (p<0.0001)
Retina Specialist 3	0.30	-0.08 (p<0.01)
Optometrist 1	0.25	-0.13 (p<0.0001)
Optometrist 2	0.27	-0.11 (p<0.0001)
Optometrist 3	0.33	-0.05 (p=0.14)

Supplementary Table 10 | Sensitivity and specificity difference between human and model in the benchmark study. Negative values show how much better the model performs over humans for either sensitivity or specificity with positive values indicating humans performing better. Below each number is the p-value obtained using McNemar's test. Bold font is used to indicate statistical significance p<0.05. The liberal model is statistically superior to each expert for sensitivity but not specificity. The conservative model is statistically superior at either specificity or/and sensitivity for each expert except Optometrist 3 who matches the conservative model the closest (p>0.05 for specificity and sensitivity). N=1053 for all values.

		Conservative model		**Liberal model**	
		Sens (0.39)	**Spec (0.89)**	**Sens (0.82)**	**Spec (0.56)**
Optometrist 1	**Sens (0.39)**	0.0 (p=1.0)		**-0.43 (p<0.0001)**	
	Spec (0.72)		**-0.17 (p<0.0001)**		+0.17 (p<0.0001)
Optometrist 2	**Sens (0.56)**	+0.17 (p<0.0001)		**-0.26 (p<0.0001)**	
	Spec (0.60)		**-0.29 (p<0.00001)**		+0.04 (p=0.014)
Optometrist 3	**Sens (0.37)**	-0.02 (p=0.76)		**-0.45 (p<0.0001)**	
	Spec (0.86)		-0.03 (p=0.11)		+0.30 (p<0.0001)
Retina Specialist 1	**Sens (0.25)**	**-0.14 (p<0.00001)**		**-0.57 (p<0.0001)**	
	Spec (0.93)		+0.04 (p<0.001)		+0.37 (p<0.0001)
Retina Specialist 2	**Sens (0.18)**	**-0.21 p<0.00001**		**-0.64 (p<0.0001)**	
	Spec (0.93)		**+0.04 (p<0.001)**		+0.37 (p<0.0001)
Retina Specialist 3	**Sens (0.29)**	**-0.10 (p=0.02)**		**-0.53 (p<0.0001)**	
	Spec (0.90)		+0.01 (p=0.32)		+0.34 (p<0.0001)

Supplementary Table 2 | Differences in demographics across sites. The dataset was collected from seven separate sites in the London area. Differences in age, sex and ethnicity are reported here for each site. For patients that attended more than one site during their history, the site that the patient attended more frequently was used for collating this data. Two sites, Darent Valley and Potters Bar have been excluded from the table - these sites had EMR data available but without imaging data and were excluded from the dataset after transfer.

Site	Total patients	% of dataset	Age		Sex		Ethnicity				
			Median age	Mean age	% female	% male	% Asian	% Black	% Other	% Unknown	% White
City Road	2092	64.0	79	77.8	59.8	40.2	8.7	2.9	13.0	10.9	64.5
Ealing	225	6.9	82	80.9	61.8	38.2	23.1	2.2	25.3	1.3	48.0
Loxford	95	2.9	83	80.9	67.4	32.6	13.7	1.1	4.2	10.5	70.5
Sir Ludwig Guttmann	45	1.4	79	77.5	57.8	42.2	8.9	2.2	8.9	8.9	71.1
Northwick Park	266	8.1	84	83.1	62.8	37.2	18.4	1.5	3.4	21.4	55.3
St. Ann's	96	2.9	81	80.6	62.5	37.5	4.2	3.1	9.4	11.5	71.9
St. George's	449	13.7	82	80.9	63.3	36.7	6.2	2.0	34.1	5.6	52.1

Supplementary Table 15 | Deformation parameters. Parameters were chosen to provide sensible deformations during augmentation of exAMD network inputs. Deformations were consistent across OCT and segmentation inputs. Each parameter was chosen carefully to not deform important imaging characteristics but help the model learn morphological invariances. Parameters are listed for train and test in the x, y, z axes. Ranges indicate uniform sampling done during each augmentation. For a description of each deformation, see https://github.com/deepmind/multidim-image-augmentation.

Deformation parameter	Train	Test
Control grid spacing (voxels)	101x101x11	101x101x11
Cropping offset (voxels)	[-16,16]x[-16,16]x[-1,1]	[-16,16]x[-16,16]x[-1,1]
Deformation magnitude (um)	7x7x0	1x1x0
Rotation (angle)	0x0x[-π/12,π/12]	0x0x0
Scaling factors	[0.8,1.2]x[0.8,1.2]x1	[0.95,1.05]x[0.95,1.05]x1
Mirror factor (probability)	0.5x0x0.5	0x0x0
Shearing coefficient	[-0.1,0.1]x[-0.1,0.1]x0	0x0x0