Author's version of De Fauw et al. "Clinically applicable deep learning for diagnosis and referral in retinal disease". *Nature Medicine* **XX**, pppp-pppp (2018), DOI: 10.1038/s41591-018-0107-6 **Under Embargo until 13 August 2018 at 1600 London time (DO NOT SHARE)**

Clinically applicable deep learning for diagnosis and referral in retinal disease

Jeffrey De FauwDawn A SimBlackwellDriesscheReena Chopra[1] [1], Joseph R Ledsam[1], Xavier Glorot[1] , Bernardino Romera-Paredes[1] , Brendan O'Donoghue [1], Faith Mackinder[11], Daniel Visentin, Stanislav Nikolov [1,3], Simon Bouton, Rosalind Raine[1] , George van den[1] , Nenad Tomasev[1] , Kareem Ayoub [3], Julian Hughes[3], Trevor [1], Sam[12],,

, Balaji Lakshminarayanan

[1], Harry Askham [1], Clemens Meyer [1], Cían O Hughes[1] [1], Geraint Rees

[2], Dominic King[1] , Alan Karthikesalingam

[2], Catherine Egan[2] , Adnan Tufail[2] , Hugh Montgomery[3] , Demis Hassabis

Back[1] , Peng T. Khaw[2], Mustafa Suleyman[1], Julien Cornebise[4,5], Pearse A. Keane[2,5]*, Olaf Ronneberger[1] [,5]*

1 DeepMind, London, UK
2 NIHR Biomedical Research Centre at Moorfields Eye Hospital and UCL Institute of Ophthalmology, London, UK
3 University College London, London, UK
4 Work performed while employed at DeepMind
5 These authors contributed equally to this work
* e-mail: pearse.keane@moorfields.nhs.uk; olafr@deepmind.com

Abstract: The volume and complexity of diagnostic imaging is increasing at a pace faster than the availability of human expertise to interpret it. Artificial intelligence has shown great promise in classifying two-dimensional photographs of some common diseases and typically relies on databases of millions of annotated images. Until now, the challenge of reaching the performance of expert clinicians in a real-world clinical pathway with three-dimensional diagnostic scans has remained unsolved. Here, we apply a novel deep learning architecture to a clinically heterogeneous set of three-dimensional optical coherence tomography (OCT) scans from patients referred to a major eye hospital. We demonstrate performance in making a referral recommendation that reaches or exceeds that of experts on a range of sight-threatening retinal diseases after training on only 15,884 scans. Moreover, we demonstrate that the tissue segmentations produced by our architecture act as a device-independent representation; referral accuracy is maintained when using tissue segmentations from a different type of device. Our work removes previous barriers to wider clinical use without prohibitive training data requirements across multiple pathologies in a real-world setting.

Introduction

Medical imaging is expanding globally at an unprecedented rate[1,2], leading to an ever-expanding quantity of data requiring human expertise and judgement to interpret and triage. In many clinical specialities there is a relative shortage of this expertise to provide timely diagnosis and referral. For example, in ophthalmology, the widespread availability of optical coherence tomography (OCT) has not been matched by the availability of expert humans to interpret scans and refer patients to the appropriate clinical care[3]. This problem is exacerbated by the dramatic increase in prevalence of sight-threatening diseases for which OCT is the gold standard of initial assessment[4–7].

Artificial intelligence (AI) provides a promising solution for such medical image interpretation and triage, but despite recent breakthroughs demonstrating expert-level performance on two-dimensional photographs in

preclinical settings[8,9], prospective clinical application of this technology remains stymied by three key challenges. First, AI (typically trained on hundreds of thousands of examples from one canonical dataset) must generalise to new populations and devices without a substantial loss of perfomance, and without prohibitive data requirements for retraining. Second, AI tools must be applicable to real-world scans, problems and pathways, and designed for clinical evaluation and deployment. Finally, AI tools must match or exceed the performance of human experts in such real-world situations. Recent work applying AI to OCT has shown promise in resolving some of these criteria in isolation, but has not yet shown clinical applicability by resolving all three.

Results

Clinical Application & AI architecture

We developed our architecture in the challenging context of optical coherence tomography (OCT) imaging for ophthalmology. We tested this approach for patient triage in a typical ophthalmology clinical referral pathway, comprising more than 50 common diagnoses for which OCT provides the definitive imaging modality (**Supplementary Table 1**) . OCT is a three-dimensional volumetric medical imaging technique analogous to 3D ultrasonography but measuring the reflection of near-infrared light rather than sound waves at a resolution for living human tissue of ~5 µm[10]. OCT is now one of the most common imaging procedures with 5.35 million OCT scans performed in the U.S. Medicare population in 2014 alone (see https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier.html). It has been widely adopted across the UK National Health Service (NHS) for comprehensive initial assessment and triage of patients requiring rapid non-elective assessment of acute and chronic sight loss. Rapid access "virtual" OCT clinics have become the standard of care [11,12]. In such clinics, expert clinicians interpret the OCT and clinical history to diagnose and triage patients with pathology affecting the macula, the central part of the retina required for high-resolution, color vision.

Automated diagnosis of a medical image, even for a single disease, faces two main challenges: technical variations in the imaging process (different devices, noise, ageing of the components, etc.), and patient-to-patient variability in pathological manifestations of disease. Existing deep learning approaches[8,9] seek to deal with all combinations of these variations using a single end-to-end black-box network, thus typically requiring millions of labeled scans. In contrast, our framework decouples the two problems (technical variations in the imaging process, and pathology variants) and solves them independently (see **Fig. 1**) . A deep segmentation network (**Fig. 1b**) creates a detailed device-independent tissue segmentation map. Subsequently, a deep classification network (**Fig. 1d**) analyses this segmentation map and provides diagnoses and referral suggestions.

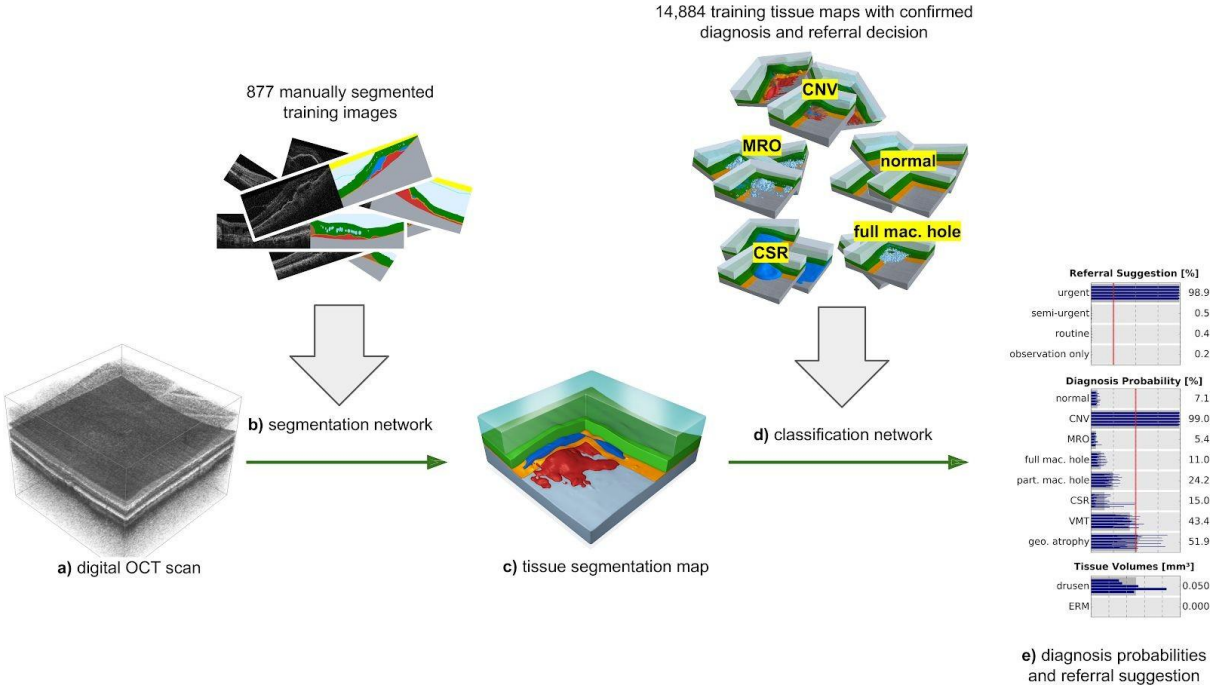


Figure 1 | Our proposed AI framework. (**a**) Raw retinal OCT scan (6 x 6 x 2.3 mm³ around the macula). (**b**) Deep segmentation network, trained with manually segmented OCT scans. (**c**) Resulting tissue segmentation map. (**d**) Deep classification network, trained with tissue maps with confirmed diagnoses and optimal referral decisions. (**e**) Predicted diagnosis probabilities and referral suggestions.

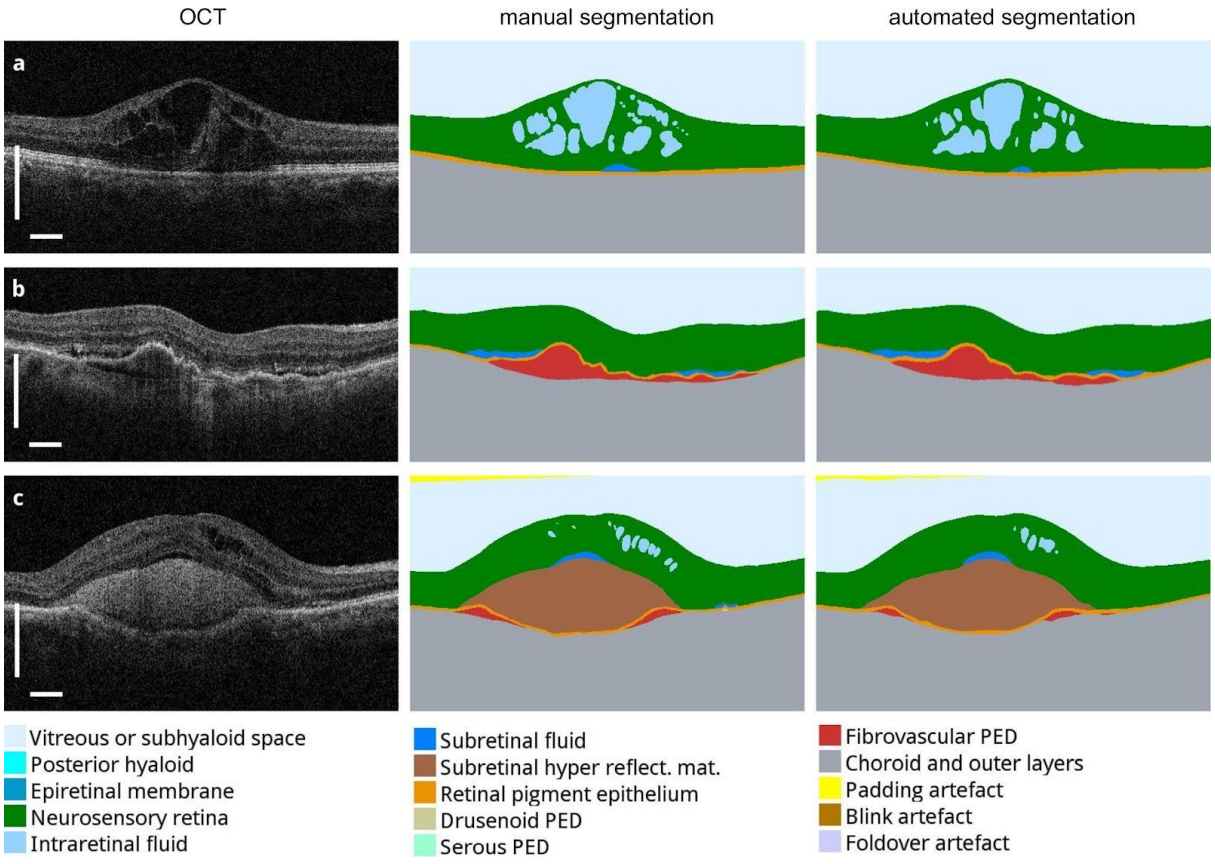


Figure 2 | Results of the segmentation network. Three selected 2D slices from the n=224 OCT scans in the segmentation test set (left column) with manual segmentation (middle column) and automated segmentation (right column; detailed color legend in **Supplementary Table 2**). (**a**) A patient with diabetic macular edema. (**b**) A patient with choroidal neovascularization resulting from age-related macular degeneration (AMD), demonstrating extensive fibrovascular pigment epithelium detachment and associated subretinal fluid. **(c)** A patient with neovascular AMD with extensive subretinal hyperreflective material. Further examples of the variation of pathology with model segmentation and diagnostic performance can be found in **Supplementary Videos 1-9**. In all examples the classification network predicted the correct diagnosis. Scale bars: 0.5mm

The segmentation network (**Fig. 1b**) uses a 3D U-Net architecture[13,14] to translate the raw OCT scan into a tissue map (**Fig. 1c**) with 15 classes including anatomy, pathology and image artefacts (**Supplementary Table 2**). It was trained with 877 clinical OCT scans (Topcon 3D OCT, Topcon, Japan) with sparse manual segmentations (Dataset #1 in **Supplementary Table 3,** see **Online Methods "Manual Segmentation" and "Datasets"** for full breakdown of scan dataset). Only approximately 3 representative slices of the 128 slices of each scan were manually segmented (see **Supplementary Table 4** for image sizes) . This sparse annotation procedure [14] allowed us to cover a large variety of scans and pathologies with the same workload as approximately 21 dense manual segmentations. Examples of the output of our segmentation network for illustrative pathologies are shown in
Fig. 2.

The classification network (**Fig. 1d**) analyses the tissue segmentation map (**Fig. 1c**) and as a primary outcome provides one of four referral suggestions currently used in clinical practice at Moorfields Eye Hospital (please

see **Supplementary Table 1** for a list of retinal conditions associated with these referral suggestions). Additionally, it reports the presence or absence of multiple, concomitant retinal pathologies (**Supplementary Table 5**). To construct the training set for this network we assembled 14,884 OCT scan volumes of 7621 patients referred to the hospital with symptoms suggestive of macular pathology (see **Online Methods "Clinical Labeling"** for details). These OCT scans were automatically segmented using our segmentation network. The resulting segmentation maps with the clinical labels built the training set for the classification network (Dataset #3 in **Supplementary Table 3,** illustrated in **Fig. 1d**).

A central challenge in OCT image segmentation is the presence of ambiguous regions, where the true tissue type cannot be deduced from the image, and thus multiple equally plausible interpretations exist. To address this issue, we trained not one but multiple instances of the segmentation network. Each network instance creates a full segmentation map for the given scan, resulting in multiple hypotheses (see **Supplementary Fig. 1**). Analogous to multiple human experts, these segmentation maps agree in areas with clear image structures but may contain different (but plausible) interpretations in ambiguous low-quality regions. These multiple segmentation hypotheses from our network can be displayed as a video, where the ambiguous regions and the proposed interpretations become clearly visible (see **Online Methods "Visualization of results in clinical practice";** use of this viewer across a range of challenging macular diseases is illustrated in **Supplementary Videos 1-9**).

Achieving Expert Performance on Referral Decisions

To evaluate our framework, we first defined a gold standard. This used information not available at the first patient visit and OCT scan, by examining the patient clinical records to determine the final diagnosis and optimal referral pathway in the light of that (subsequently obtained) information. Such a gold standard can only be obtained retrospectively. Gold standard labels were acquired for 997 patients not included in the training dataset (Dataset #5 in **Supplementary Table 5).** We then tested our framework on this dataset. For each patient, we obtained the referral suggestion of our framework plus an independent referral suggestion from eight clinical experts, four of whom were retina specialists and four optometrists trained in medical retina; see **Supplementary Table 6** for more information. Each expert provided two separate decisions, one (like our framework) from the OCT scan alone (Dataset #7 in **Supplementary Table 5**); and one from the OCT plus fundus image and clinical notes (Dataset #8 in **Supplementary Table 5,** see **Supplementary Fig. 2**), in two separate sessions spaced at least two weeks apart. We compared each of these performances (framework and two expert decisions) against the gold standard.

Our framework achieved and in some cases exceeded expert performance (**Fig. 3**). To illustrate this, **Fig. 3a** displays performance on "Urgent referrals", the most important clinical referral decision (mainly due to pathologies that cause choroidal neovascularization (CNV) -- see **Supplementary Table 1**) versus all other referral decisions as a receiver operating characteristic (ROC) plot (plots for the other decisions are shown in **Supplementary Fig. 3**). Performance of our framework matched our two best retina specialists and had a significantly higher performance than the other two retinal specialists and all four optometrists when they used only the OCT scans to make their referral suggestion. (**Fig. 3a** filled markers). When experts had access to the fundus image and patient summary notes to make their decision, their performance improved (**Fig. 3a** empty markers) but our framework remained as good as the five best experts and continued to significantly outperform the other three (see Supplemental Material).

To provide a fuller picture, the overall performance of our framework on all four clinical referral suggestions ("urgent", "semi-urgent", "routine", and "observation only") is displayed in **Fig. 3b** compared to the two highest performing retina specialists. The framework performed comparably to the two best-performing retina specialists, and made no clinically serious wrong decisions (topright element of each matrix, i.e. referring a patient who needs an urgent referral to observation only). Confusion matrices for the assessments of the other human experts are shown in **Supplementary Fig. 4**. The aggregated number of wrong referral decisions is displayed as error rate (1 - accuracy) for our framework and all experts in **Fig. 3c** . Our framework

(5.5% error rate) performed comparably to the two best retina specialists (6.7% and 6.8% error rate) and significantly outperformed the other six experts in the "OCT only" setting. Significance thresholds (3.9% for higher performance and 7.3% for lower performance) were derived by a two-sided exact binomial test, incorporating uncertainty both from expert and from algorithm (see **Online Methods "Statistical Analysis"**). When experts additionally used the fundus image and the patient's summary notes, five approached the performance of our framework (three retina specialists and two optometrists), which continued to significantly outperform the remaining three (one retina specialist and two optometrists).

Our framework uses an ensemble of five segmentation and five classification model instances (see **Supplementary Fig. 1**) to achieve these results. Beside the benefits of an uncertainty measure, ensembling also significantly improves overall performance compared to a single model instance. Error rates for different ensemble sizes are shown in **Supplementary Fig. 5.** With more segmentation model instances and more classification model instances, performance increases. The bottom right cells in that table illustrate that performance differences between 4 x 4 model instances and 5 x 5 model instances are only marginal, so we do not expect significant changes by adding more instances.

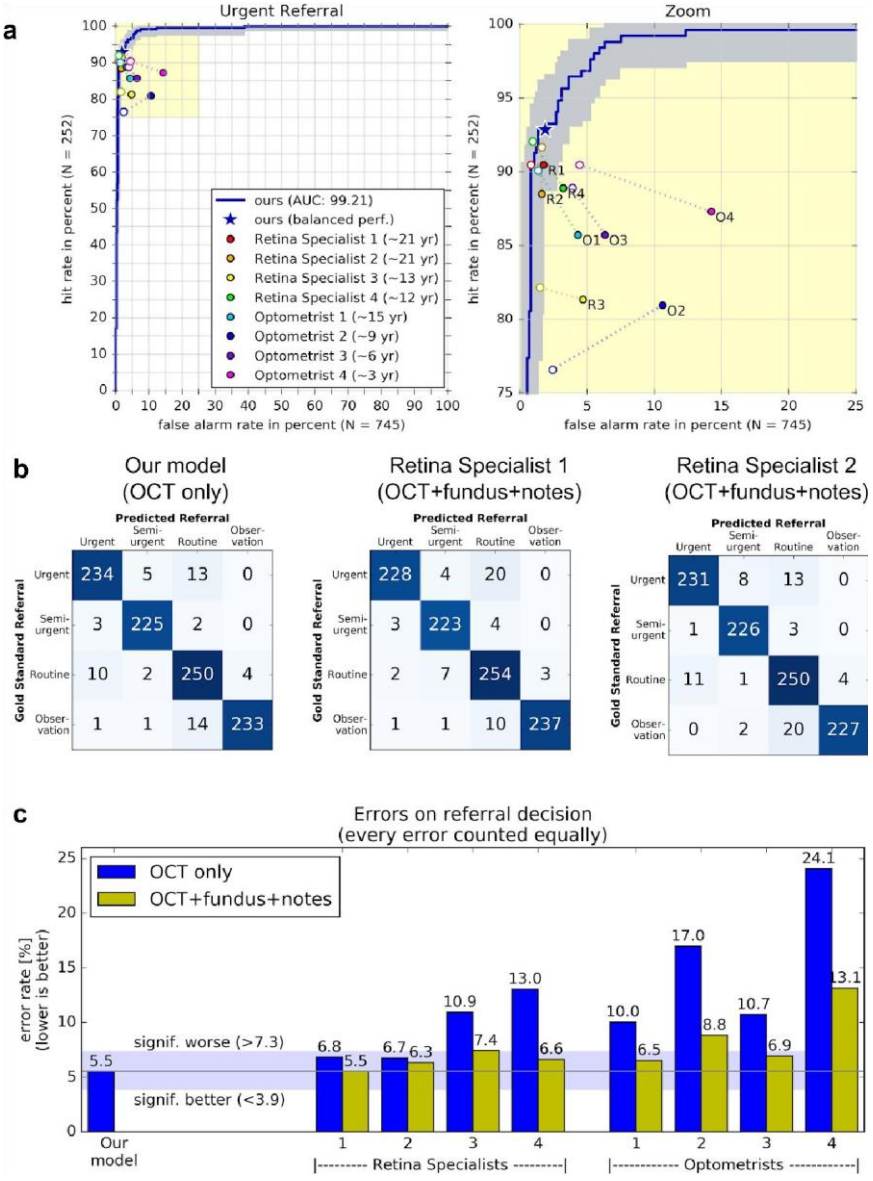


Figure 3 | Results on the patient referral decision. Performance on an independent test set of n=997 patients (252 urgent, 230 semi-urgent, 266 routine, 249 observation only). (**a**) Receiver operating characteristic (ROC) diagram for "urgent referral" (due to choroidal neovascularization (CNV)) versus all other referrals. The blue ROC curve is created by sweeping a threshold over the predicted probability of a particular clinical diagnosis. Points outside the light blue area correspond to a significantly different performance (95% confidence level, using a two-sided exact binomial test). The asterisk denotes the performance of our model in the 'balanced performance' setting. Filled markers denote experts' performance using OCT only; emptyoutlined markers denote their performance using OCT, fundus image and summary notes. Dashed lines connect the two performance

points of each expert. (**b**) Confusion matrices with patient numbers for referral decision for our framework and the two best retina specialists. These show the number of patients for each combination of gold standard decision and predicted decision. The numbers of correct decisions are found on the diagonal. Wrong decisions due to overdiagnosis are in the lower-left triangle, and wrong decisions due to underdiagnosis are in the upper-right triangle. (**c**) Total error rate (1 - accuracy) on referral decision. Values outside the light blue area (3.9% - 7.3%) are significantly different (95% confidence interval, using a two-sided exact binomial test) to the framework performance (5.5%). AUC: area under curve.

The accumulated number of diagnostic errors does not fully reflect the clinical consequences that an incorrect referral decision might have for patients, which depends also on the specific diagnosis missed. For example, failing to diagnose sight-threatening conditions could result in rapid visual loss[3,15,16] which is not the case for many other diagnoses. For an initial quantitative estimation of these consequences, we weighted different types of diagnostic errors according to our clinical experts' judgement of the clinical impact of erroneous classification (expressed as penalty points; see **Supplementary Fig. 6a**) . We derived a score for our framework and each expert as a weighted average of all wrong diagnoses. This revealed that our framework achieved a lower average penalty point score than any of our experts (**Supplementary Fig. 6b**) . We further optimized our framework decisions to minimise this specific score (see **Online Methods "Optimizing the Ensemble Output for Sensitivity, Specificity and Penalty Scores"**) which further improved performance (**Supplementary Fig. 6b**). Thus the expert performance of our framework is not achieved at the cost of missing clinically important sight-threatening diagnoses.

To examine how our proposed two-stage architecture compared to a traditional single-stage architecture, we trained an end-to-end classification network with the same architecture as our second stage to map directly from a raw OCT scan to a referral decision (see **Methods "End-to-end Classification Network"**). The error rate achieved with an ensemble of five network instances was 5.5%, which was not significantly different from the performance of the two-stage architecture. This validates our choice of the two-stage architecture that offers several clinical advantages. See **Supplementary Fig. 7** for detailed results.

Achieving Expert Performance on Retinal Morphology

The referral decision recommended by our framework is determined by the most urgent diagnosis detected on each scan (**Supplementary Table 1**). Patients may also have multiple concomitant retinal pathologies. These additional pathologies do not change the referral decision, but may have implications for further investigations and treatment. Our framework was therefore also trained to predict the probability of a patient having one or more of several pathologies (**Supplementary Table 5**).

To evaluate performance on diagnosing multiple pathologies, a 'silver standard' for each scan was established by majority vote from eight experts who evaluated the OCT scan, fundus image and patient summary notes (Dataset #6 in **Supplementary Table 3)**. This majority vote biases the assessment against our framework. Nevertheless, our framework demonstrated an area under the ROC curve that was over 99% for most of the pathologies (and over 96% for all of them; **Supplementary Table 7**), on par with the experts' performance on OCT only. As with earlier evaluations, experts' performance improved when they were provided also with the fundus image and patient summary notes. This improvement was most marked in pathologies classed as 'routine referral' e.g. geographic atrophy and central serous retinopathy. Many of these are conditions where the fundus photograph or demographic information would be expected to provide important information, indicating that there is scope for future work to improve the model. However even in the worst case our framework still performed on par with at least one retinal specialist and one optometrist (**Supplementary Table 6** and **Supplementary Fig. 8**).

Generalization to a New Scanning Device Type

A key benefit of our two-stage framework is the device independence of the second stage. Using our framework on a new device generation thus only requires retraining of the segmentation stage to learn how each tissue type appears in the new scan, while knowledge about patient-to-patient variability in pathological manifestation of different diseases that it learned from the approximately 15,000 training cases can be reused. To demonstrate this generalization, we collected an independent test set of clinical scans from 116 patients (plus confirmed clinical outcomes) recorded with a different OCT scanner type from a different vendor (Spectralis, Heidelberg

Engineering, Germany; hereafter "device type 2") . This dataset is listed as Dataset #11 in **Supplementary Table 3** (see also **Online Methods "Datasets"** for details). We selected this device type for several reasons. It is the second most used device type at Moorfields Eye hospital for these examinations, giving rise to a sufficient number of scans. It has a similar worldwide market share as device type 1. But most importantly, this device type provides a large difference in scan characteristics compared to the original device type (see **Supplementary Fig. 9**).

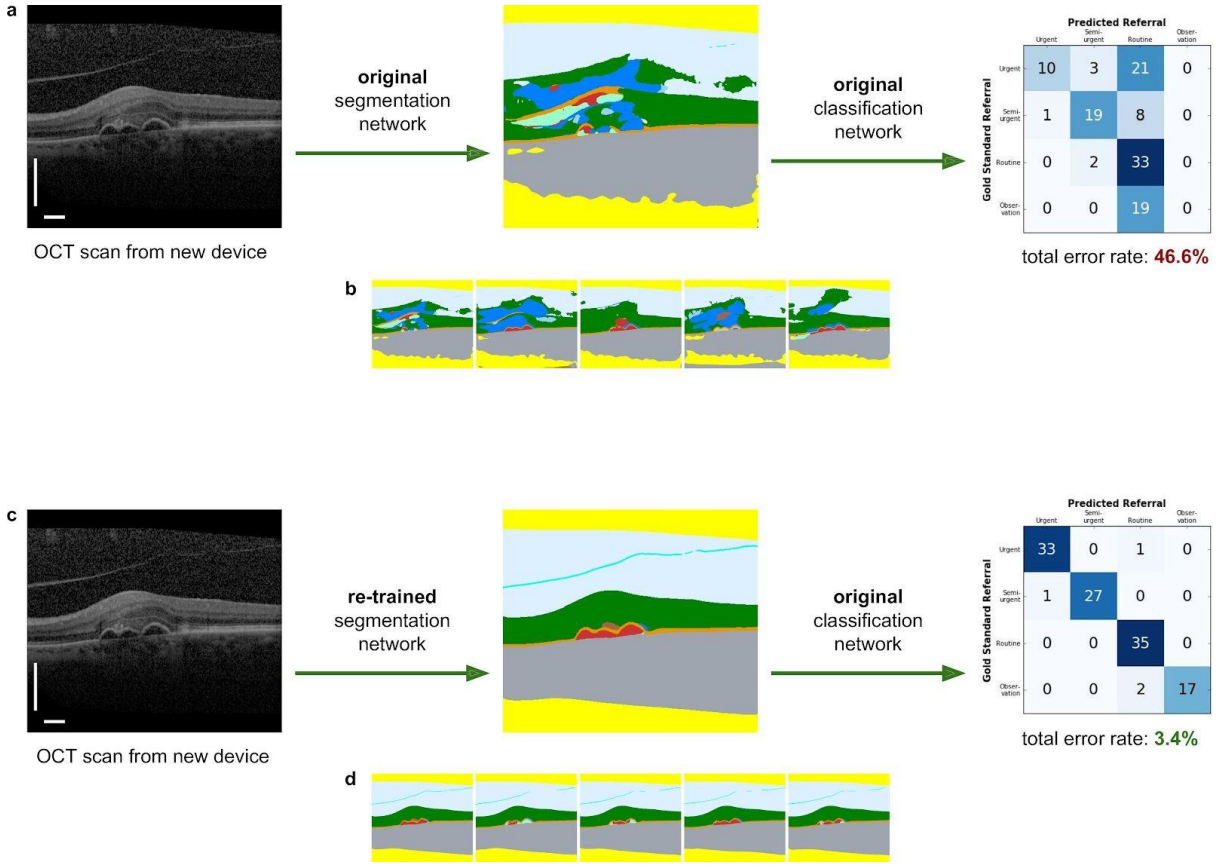


Figure 4 | Generalization to a new scanning device type. (**a**) Low performance of original network on OCT scans from the new device type 2. Left: The selected slice shows the different appearance of structures in device type 2. Middle: A poor quality segmentation map created with our original segmentation network (color legend in **Supplementary Table 2**). Right: Resulting performance on a new test set of n=116 patients. The confusion matrix shows patient numbers for the referral suggestion. (**b**) All five segmentation hypotheses from our original network. The strong variations show the large uncertainty. (**c**) High performance was attained on the device type 2 test set (n=116) after re-training the segmentation network with OCT scans from device type 1 and device type 2. The classification network is unchanged. (**d**) All five segmentation hypotheses from the re-trained segmentation network. The network is confident in the interpretation of most structures, and just highlights the ambiguities in the sub-retinal pigment epithelium (RPE) space. Scale bars: 0.5mm

To evaluate the effect of a different scanning device type, we initially fed the OCT scans from device type 2 into our framework trained only on scans from device type 1 (**Fig. 4a**). The segmentation network is clearly confused by the changed appearance of these structures and attempted to explain them as additional retinal layers (**Fig. 4a** middle). Consequently, performance was poor with a total error rate for referral suggestions of 46.6% (**Fig. 4a** right). Uncertainty of the segmentation network on these (never seen) types of images resulted in five strongly different segmentation hypotheses (**Fig. 4b**).

We next collected an additional segmentation training set with 152 scans (527 manually segmented slices in total) from this device (Dataset #9 in **Supplementary Table 3)**, and retrained the segmentation network with

both the training scans from the original device type 1 and the new device type 2 (see **Online Methods "Segmentation Network"** for details). The classification network was not modified.

Our retrained system (adapted segmentation network + unchanged classification network) now achieved a similarly high level of performance on device type 2 as on the original device (**Fig. 4c**). It suggested incorrect referral decisions in 4 of the 116 cases, a total error rate of 3.4%. Due to the small number of cases in the new test set, this is not significantly different to the error rate of 5.5% on device type 1 (P(4 out of 116 < 55 out of 997) = 0.774, see **Online Methods "Statistical Analysis"**). For continuity with our previous evaluation, we also measured performance against retina specialists accessing OCT scans plus fundus images and clinical notes (Dataset #12 in **Supplementary Table 3**). Our experts achieved the following error rates (all with access to imaging and clinical notes): retinal specialist one: 2 errors = 1.7% error rate; retinal specialist two: 2 errors = 1.7% error rate; retinal specialist three: 4 errors = 3.4% error rate; retinal specialist four: 3 errors = 2.6% error rate; retinal specialist five: 3 errors = 2.6% error rate. These differences in performance between our framework and the best human retina specialists did not reach statistical significance (P(4 out of 116 > 2 out of 116) = 0.776).

To verify that device type 2 provides the greatest difference in scan characteristics, we performed a feasibility study on the small number of OCT scans from Cirrus HD-OCT 5000 with AngioPlex (Carl Zeiss Meditec) devices available in Moorfields Eye Hospital (dataset of 61 scans not included here). Applying our original network to these images we already obtained an error rate of 16.4%. This rate was much lower than that originally obtained with device type 2 (46.6%), consistent with the claim that device type 2 provides a larger difference in scan characteristics from device type 1. Retraining of the segmentation network with 6 manually segmented scans reduced the error rate to 9.8%.

Table 1 summarizes our results: For device type 1 our architecture required 877 training scans with manual segmentations and 14,884 training scans with gold standard referral decisions to achieve expert performance on referral decisions (5.5% error rate). For device type 2 we only required 152 additional training scans with manual segmentations and not a single additional training scan with gold standard referral decisions to achieve the same performance on referral decisions on this device type (3.4% error rate).

Table 1 | Number of training scans and achieved performance on the two device types

	Training Scans with sparse manual segmentations	**Training** Scans with gold standard referral decision	**Test** Performance on referral decision (error rate)	**Test** Performance on urgent referral (AUC)
Device type 1	877	14,884	55 out of 997 = **5.5%**	99.21
Device type 2	152 (+ 877 scans from device type 1)	0	4 out of 116 = **3.4%**	99.93

Discussion

Recent work applying AI to the automated diagnosis of OCT scans shows encouraging results but until now such studies have relied on selective and clinically unrepresentative OCT datasets. For example, several authors [17–21] report high performance on automated classification of age-related macular degeneration (AMD) from OCT scans. However, they tested their algorithms on smaller datasets that exclude other pathologies. In contrast, here we demonstrate expert performance on multiple clinical referral suggestions for two independent test datasets of 997 and 116 clinical OCT scans that include a wide range of retinal pathologies.

Several recent studies used deep learning based architectures to deliver successful segmentation of OCT scans[22–25]. This earlier work focused on a subset of diagnostically relevant tissues types (e.g. intraretinal fluid) and applied 2D models in samples of between 10 and 42 patients. In the present work we go beyond these earlier studies by applying 3D models, segmenting a much larger range of diagnostically relevant tissue types, and connect such segmentation to clinically relevant real-world referral recommendations.

We evaluated our framework on a broad range of real-world images from routine clinical practice at 32 different Moorfields Eye Hospital sites covering diverse populations within London and surrounding areas, using 37 individual OCT devices (28 device type 1 and 9 device type 2). The two device types we tested are both used widely in routine clinical practice at Moorfields Eye Hospital, the largest eye hospital in Europe and North America, and provided a large difference in scan characteristics.

A number of potential benefits extend from our framework. The derivation of a device-independent segmentation of the OCT scan creates an intermediate representation that is readily viewable by a clinical expert and integrates into clinical workflows (see **Fig. 5** for the clinical results viewer). Moreover, the use of an ensemble of five segmentation network instances allows us to present ambiguities arising from the imaging process to the decision network (and could potentially be used for automated quality control).

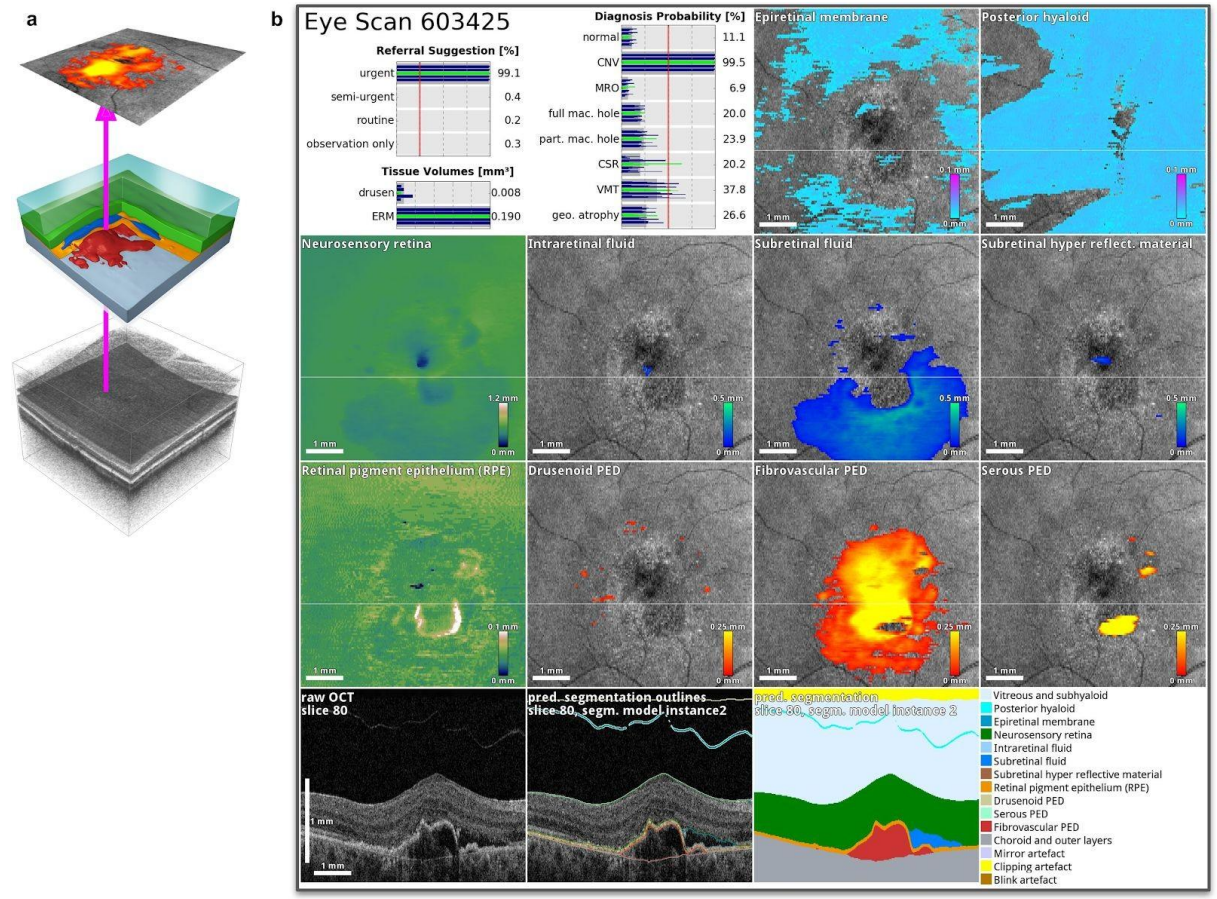


Figure 5 | Visualization of the segmentation results as thickness maps. (**a**) The average intensity projection of the OCT scan along A-scan direction (frontal view of the eye) is overlaid with a thickness map of the fibrovascular pigment epithelium detachment (PED,red segment). (**b**) Screenshot from our OCT viewer. (**Row 1 left**) Referral suggestion, tissue volumes and diagnosis probabilities. The highlighted bars correspond to the selected segmentation model. (**Rows 1-3)** Thickness maps of the 10 relevant tissue types from segmentation model instance 2. The two healthy tissue types (high level retina and RPE) are displayed in a black-blue-green-brown-white color map, the pathological tissues (all others) are displayed as overlay on a projection of the raw OCT scan. The thin white line indicates the position of slice 80. (**Row 4**) Slice 80 from the OCT scan and the segmentation map from segmentation model instance 2. Detailed tissue legend in **Supplementary Table 2** . The slice and model instance can be interactively selected (see **Supplementary Video 1**).

The 'black box' problem has been identified as an impediment to the application of deep learning in healthcare[2,6]. Here we created a framework whose structure closely matches the clinical decision-making process, separating judgements about the scan itself from the subsequent referral decision. This allows a clinician to

inspect and visualize an interpretable segmentation, rather than simply being presented with a diagnosis and referral suggestion. Such an approach to medical imaging AI offers potential insights into the decision process, in a fashion more typical of clinical practice. For example, an interpretable representation is particularly useful in difficult and ambiguous cases. Such cases are common in medicine and even expert medical practitioners can find it difficult to reach consensus (for example, our eight experts only agreed on 63.5% of cases even when accessing all information).

Our segmentation map assigns only one label per pixel, and it may not be possible to use the framework directly in other clinical pathways where the tissue segmentation map does not contain all required information for a diagnosis (e.g. in certain radiomics applications). To keep the advantages of the intermediate device-independent representation in such applications, future work can potentially augment the tissue segmentation map with multiple labels per pixel to encode local tissue features, or with additional channels that encode continuous features like inflammatory reaction. This may be of particular value for other components of the retina such as the nerve fibre layer, and may be of importance for multiple ocular and brain disorders such as glaucoma and dementia.

While we have demonstrated the performance of our framework in the domain of a clinical treatment pathway, the approach has potential utility in clinical training where medical professionals must learn to read medical images. In addition, a wide variety of non-medically qualified health professionals have an interest in appropriately reading and understanding medical images. Our framework produces a visualisable segmentation and achieves expert performance on diagnosis and referral decisions for a large number of scans and pathologies. This therefore raises the intriguing possibility that such a framework could be evaluated as a tool for effectively training health care professionals to expert levels.

Segmentation output itself can also be used to quantify retinal morphology and derive measurements of particular pathologies (for example, the location and volume of fibrovascular pigment epithelium detachment and macular edema). Some of these measurements (such as retinal thickness and intraretinal fluid) can currently be derived automaticallyclinical trials of therapies for retinal disease[27,28], used to investigate correlations with visual outcomes[29–32]. Our framework can be used to define and validate a broader[27] and as an endpoint in range of automatically derived quantitative measurements.

Our framework can triage scans at first presentation of a patient into a small number of pathways used in routine clinical practice with a performance matching or exceeding both expert retina specialists and optometrists who staff virtual clinics in a UK NHS setting. Future work can now directly seek evidence for efficacy of such a framework in a randomized controlled trial. The output of our framework can be optimized to penalize different diagnostic errors, and thus for other clinically important metrics. For example, the potential improvement to patient quality of life of different diagnostic decisions, or avoiding the harm of unnecessary investigation that might come from a false-positive diagnosis, could all be incorporated into future work.

Globally, ophthalmology clinical referral pathways vary, and the range of diseases that can potentially be diagnosed by OCT includes pathologies additional to those macular diseases studied here. We studied a major clinical referral pathway in a global center of clinical excellence focusing on 53 key diagnoses relevant to the national (NHS) referral pathways. Our work opens up the possibility of testing the clinical applicability of this approach in other global settings and clinical pathways such as emergency macular assessment clinics in the UK NHS, triage and assessment in community eye care centers and the monitoring of disease during treatment regimes. Furthermore, devices such as binocular OCT[33] have the potential to increase accessibility in emerging economies. Images produced by such devices will differ in resolution, contrast and image quality from the state-of-the-art devices studied here, and existing AI models trained on current state-of-the-art devices may perform poorly on such new devices. Our proposed two-stage model offers exciting possibilities in deploying models more efficiently in countries where state-of-the-art OCT devices are too costly for widespread adoption.

In conclusion, we present a novel framework that analyses clinical OCT scans and makes referral suggestions to a standard comparable to clinical experts. While focused on one common type of medical imaging, future work

can address a much wider range of medical imaging techniques, and incorporate clinical diagnoses and tissue types well outside the immediate application demonstrated here.

References

1. OECD. Computed tomography (CT) exams (indicator). (2017). doi:10.1787/3c994537-en
2. OECD. Magnetic resonance imaging (MRI) exams (indicator). (2017). doi:10.1787/1d89353f-en
3. Foot, B. & MacEwen, C. Surveillance of sight loss due to delay in ophthalmic treatment or review: frequency, cause and outcome. *Eye* **31,** 771–775 (2017).
4. 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).
5. Rudnicka, A. R. *et al.* Incidence of late-stage age-related macular degeneration in American whites: systematic review and meta-analysis. *Am. J. Ophthalmol.* **160,** 85–93 (2015).
6. Bourne, R. R. A. *et al.* Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis. *T he Lancet Global Health* **5,** e888–e897 (2017).
7. Schmidt-Erfurth, U., Klimscha, S., Waldstein, S. M. & Bogunović, H. A view of the current and future role of optical coherence tomography in the management of age-related macular degeneration. *E ye* **31,** 26–44 (2017).
8. Gulshan, V. *et al.* Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316,** 2402–2410 (2016).
9. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542,** 115 (2017).
10. Huang, D. *et al.* Optical coherence tomography. *Science* **254,** 1178–1181 (1991).
11. Buchan, J. C. *e t al.* How to defuse a demographic time bomb: the way forward? *E ye* **31 ,** 1519–1522 (2017).
12. Whited, J. D. *et al.* A modeled economic analysis of a digital teleophthalmology system as used by three federal healthcare agencies for detecting proliferative diabetic retinopathy. *T elemedicine and e-Health* **11 ,** 641–651 (2005).
13. Ronneberger, O., Fischer, P. & Brox, T. U-Net: convolutional networks for biomedical image segmentation. in *Med Image Comput Comput Assist Interv* 234–241 (Springer International Publishing, 2015).
14. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. *Med Image Comput Comput Assist Interv* 424–432 (2016).
15. Muether, P. S., Hermann, M. M., Koch, K. & Fauser, S. Delay between medical indication to anti-VEGF treatment in age-related macular degeneration can result in a loss of visual acuity. *G raefes Arch. Clin. Exp. Ophthalmol.* **249,** 633–637 (2011).
16. Arias, L. *et al.* Delay in treating age-related macular degeneration in Spain is associated with progressive vision loss. *Eye* **23,** 326–333 (2009).
17. Karri, S. P. K., Chakraborty, D. & Chatterjee, J. Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration. *B iomed. Opt. Express* **8,** 579 (2017).
18. Apostolopoulos, S., Ciller, C., De Zanet, S. I., Wolf, S. & Sznitman, R. RetiNet: automatic AMD identification in OCT volumetric data. *Preprint at http://arxiv.org/abs/1610.03628v1* (2016).

19. Farsiu, S. *et al.* Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography. *Ophthalmology* **1 21,** 162–172 (2014).
20. Srinivasan, P. P. *et al.* Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images. *Biomed. Opt. Expres s* **5 ,** 3568–3577 (2014) .
21. Lee, C. S., Baughman, D. M. & Lee, A. Y. Deep learning is effective for classifying normal versus age-related macular degeneration OCT images. *Ophthalmol Retina* **1,** 322–327 (2017) .
22. Fang, L. *et al.* Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search. *Biomed. Opt. Express* **8,** 2732–2744 (2017).
23. Lee, C. S. *et al* . Deep-learning based, automated segmentation of macular edema In optical coherence tomography. *Biomed. Opt. Expres s* **8,** 3440–3448 (2017).
24. Lu, D. *et al.* Retinal fluid segmentation and detection in optical coherence tomography images using fully convolutional neural network. *P reprint at http://arxiv.org/abs/1710.04778v1* (2017).
25. Roy, A. G. *et al.* ReLayNet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional network. *Biomed. Opt. Express* **8,** 3627–3642 (2017).
26. Castelvecchi, D. Can we open the black box of AI? *Nature News* **538 ,** 20 (2016).
27. Schmidt-Erfurth, U. *et al.* Machine learning to analyze the prognostic value of current imaging biomarkers in neovascular age-related macular degeneration. *O phthalmol Retina* **2 ,** 24–30 (2018) .
28. Schlegl, T. *et al.* Fully automated detection and quantification of macular fluid in OCT using deep learning. *Ophthalmology* **125,** 549–558 (2018).
29. Keane, P. A. & Sadda, S. R. Predicting visual outcomes for macular disease using optical coherence tomography. *Saudi J Ophthalmol* **25,** 145–158 (2011).
30. Schaal, K. B., Rosenfeld, P. J., Gregori, G., Yehoshua, Z. & Feuer, W. J. Anatomic clinical trial endpoints for nonexudative age-related macular degeneration. *Ophthalmology* **123,** 1060–1079 (2016).
31. Schmidt-Erfurth, U. & Waldstein, S. M. A paradigm shift in imaging biomarkers in neovascular age-related macular degeneration. *P rog. Retin. Eye Res.* **50 ,** 1–24 (2016).
32. Villani, E. *et al.* Decade-long profile of imaging biomarker use in ophthalmic clinical trials. *Invest. Ophthalmol. Vis. Sci.* **58,** BIO76–BIO81 (2017).
33. Chopra, R., Mulholland, P. J., Dubis, A. M., Anderson, R. S. & Keane, P. A. Human factor and usability testing of a binocular optical coherence tomography system. *Transl. Vis. Sci. Technol.* **6,** 16 (2017).

Acknowledgements

We thank K. Kavukcuoglu, A. Zisserman, M. Jaderberg, K. Simonyan for discussions, A. Cain and M. Cant for work on the visuals, D. Mitchell and M. Johnson for infrastructure and systems administration, J. Morgan and OpenEyes for providing the EHR records, T. Peto, P. Blows, A. O'Shea, and the NIHR Clinical Research Facility for work on the labeling, T. Heeran, M. Lukic, K. Kortum, K. Fasler, S. Wagner and N. Pontikos for work on the labeling, E. Steele, V. Louw, S. Gill, and the rest of Moorfields IT team for work on the data collection and de-identification, S. Al-Abed and N. Smith for Moorfields technical advice at project initiation, R. Wood and D. Corder at Softwire for engineering support at Moorfields, R. Ogbe and the Moorfields Information Governance team for support, M. Hassard for Moorfields research & development support, K. Bonstein and the National Institute for Health Research (NIHR) for support at the Moorfields Biomedical Research Centre (BRC), J. Besley for legal assistance, E. Manna for patient engagement and support, and the rest of the DeepMind team for their support, ideas and encouragement.

Dr Keane is supported by an NIHR Clinician Scientist Award (NIHR-CS--2014-14-023). Drs Sim, Tufail and Egan, and Prof Sir Khaw are supported by the NIHR Biomedical Research Centre at Moorfields Eye Hospital

NHS Foundation Trust and UCL Institute of Ophthalmology and the NIHR Moorfields Clinical Research Facility. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. Ms. Chopra receives studentship support from the College of Optometrists, UK.

Author contributions

P.A.K., M.S., J.C., D.H., P.T.K, T.B., K.A. initiated the project and the collaboration.
O.R., J.D.F., B.R.-P., S.N. developed the network architectures, training and testing setup.
P.A.K., J.R.L., R.C. designed the clinical setup.
P.A.K., J.R.L., J.C., R.C., D.A.S., C.E., A.T. created the data set and defined clinical labels.
J.D.F., B.R.-P., S.N., N.T., S.Bl., H.A., B.O'D., D.V., G.V.D.D., O.R., J.C. contributed to the software engineering.
J.R.L., S.Bl., H.A., created the database.
P.A.K., J.R.L., D.K., A.K., C.O.H., R.R., contributed clinical expertise.
O.R., P.A.K., J.D.F., J.R.L., B.R.-P., S.N., N.T., X.G. analysed the data.
T.B., S.Bou., J.C., J.H., F.M., C.M. managed the project.
O.R., P.A.K., J.R.L., J.D.F., B.R.-P., G.R., H.M. wrote the paper.
B.L. contributed to the uncertainty estimation.

Competing financial interests

P.A.K, G.R., H.M. and R.R. are paid contractors of DeepMind. Dr. Keane has received speaker fees from Heidelberg Engineering, Topcon, Haag-Streit, Allergan, Novartis, and Bayer. Dr. Keane has served on advisory boards for Novartis and Bayer, and is an external consultant for DeepMind and Optos. Professor Tufail has served on Advisory Boards for the following companies: Allergan, Bayer, Genentech, GlaxoSmithKline, Novartis, Roche. Ms Egan has received speaker fees from Heidelberg Engineering and Haag-Streit UK. Professor Sir Peng Khaw has served on advisory boards for Aerie, Allergan, Alcon, Belkin laser, Novartis, and Santen. Ms. Sim has received speaker fees from Novartis, Bayer, Allergan, Haag-Streit. The authors have no other competing interests to disclose.

Online Methods

Ethics and Information governance

This work, and the collection of 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 the Information Commissioner's Anonymization: managing data protection risk code of practice (https://ico.org.uk/media/1061/anonymisation-code.pdf), 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[43].

Visualization of Results in Clinical Practice

To facilitate viewing of the results in routine clinical practice, we display the obtained three dimensional segmentation maps as two-dimensional thickness maps overlaid on a projection of the raw OCT scan (**Fig. 5a**). The thickness maps for all tissue types are displayed side-by-side in our interactive OCT viewer (**Fig. 5b** and **Supplementary Video 1**). Our system also provides measures for its degree of certainty on both overall referral decision, and each specific retinal disease feature. In most common clinical scenarios, the algorithm will both provide the diagnosis with a high degree of certainty and highlight classical disease features (e.g. "wet" AMD -

Supplementary Video 2) . This visualization may be particularly useful in difficult and ambiguous cases, such as the diagnosis of CNV formation in cases of chronic central serous retinopathy (CSR, **Supplementary Videos 5** and **7**) or in advanced geographic atrophy due to AMD (**Supplementary Video 6**). Such a visualization may also allow clinicians to discard an automated diagnosis or referral suggestion in obvious failure cases, such as when poor image quality leads to erroneous segmentation results (**Supplementary Video 8**). Furthermore, in a screening context the tissue segmentation map can facilitate quality assurance procedures, whether that be in normal cases (**Supplementary Video 3**) or in disease cases (e.g., diabetic macular edema in the context of diabetic retinopathy screening, **Supplementary Video 4**).

Datasets and Clinical Taxonomy

Datasets

Data were selected from a retrospective cohort of all patients attending Moorfields Eye Hospital NHS Foundation Trust, a world renowned tertiary referral center with 32 clinic sites serving an urban, mixed socioeconomic and ethnicity population centered around London, U.K., between 1 June 2012 and 31 January 2017, who had OCT imaging (Topcon 3D OCT, Topcon, Japan; Spectralis, Heidelberg Engineering, Germany) as part of their routine clinical care. Conditions with fewer than ten cases, and data from patients who had manually requested that their data should not be shared, were excluded before research began. OCT scan sets containing severe artefacts, or significant reductions in signal strength to the point where retinal interfaces could not be identified were also excluded from the study (**Supplementary Fig. 10**), as such scans are non-diagnostic and in practice would usually be retaken. Scans where no diagnostic label could be attached (as described below) were excluded from the present study. For OCT examinations labeled as urgent or semi-urgent in the Moorfields OpenEyes electronic health record (EHR) only scans taken prior to treatment beginning were included; during treatment, resolution of pathology invalidates the database labels. The dataset selection and stratification process is displayed in a CONSORT flow diagram in **Supplementary Fig. 11**.

Two OCT device types were selected for investigation. 3D OCT-2000 (Topcon, Japan) was selected as “device type 1” due to its routine use in the clinical pathway we studied. For device type 1, a total of 15,877 OCT scans from 7981 individual patients (mean age 69.5; 3686 male, 4294 female, 1 gender unknown) were eligible for inclusion in the work (Datasets #3 + #4 in **Supplementary Table 3**). To create a test set representative of the real-world clinical application, 997 additional patients (mean age 63.1; 443 male, 551 female, 3 gender unknown) presenting to Moorfields with visual disturbance during the retrospective period were selected and only their referral OCT examination was selected for inclusion in the test set (Dataset #5 in **Supplementary Table 3**); a sample size requirement of 553 to detect sensitivity and specificity at 0.05 marginal error and 95% confidence was used to inform the number included. To demonstrate the generalizability of our approach, Spectralis OCT (Heidelberg Engineering, Germany) was chosen as “device type 2”. For generalisability experiments, a second test set of clinical OCT scans from 116 patients (mean age 58.2; 59 male, 57 female) presenting in the same manner were selected using the same methodology and selection criteria (Dataset #11 in **Supplementary Table 3**). Examples of differences between the two devices types are shown in **Supplementary Fig. 9. Supplementary Table 8** shows a breakdown of patients and triage categories in the datasets.

Clinical Taxonomy

OCT examinations were mapped from individual diagnoses and treatment information to specific triage decisions ("urgent referral", "semi-urgent referral", "routine referral", and "observation only") to a medical retina clinic setting (**Supplementary Table 1)**. Where possible the presence or absence of additional pathologies was added as a label (**Supplementary Table 5**). The dataset represents the full variety of medical retina patients presenting and receiving treatment at Moorfields Eye Hospital. Although the exact mapping was chosen to be relevant to the triage decisions at Moorfields Eye Hospital where the research work took place, the framework is generalisable to other systems at centers with different triage requirements (e.g., optometrists working in a high street clinic setting or ophthalmologists without subspecialty retinal expertise). Scans meeting the exclusion criteria were removed from the database before splitting the data into training, validation and test sets. **Supplementary Fig. 12** provides an example of variation within the ‘urgent referral’ label class.

Clinical Labeling

Clinical labels for the 14,884 scans in Dataset #3 in **Supplementary Table 3** were assigned through an automated notes search with trained ophthalmologist and optometrist review of the OCT scans. The presence or absence of choroidal neovascularization, referable macular edema, normal and other pathologies visible on the OCT scan were recorded. In addition, patients with choroidal neovascularization or macular edema confirmed through treatment were labeled directly from the Moorfields OpenEyes electronic health record (EHR). A validation subset of 993 scans (993 patients) was graded separately by three junior graders (ophthalmologists specializing in medical retina) with disagreement in clinical labels arbitrated by a senior retinal specialist with over 10 years experience and image reading center certification for OCT segmentation (Dataset #4 in **Supplementary Table 3**). The test set was further verified by full notes review with access to follow up data with both junior and senior grader review. Junior and senior graders were separate to those participating in the evaluation of expert performance.

Manual Segmentation

A subset of 1101 scans from device type 1 and a set of 264 scans from device type 2 were manually segmented using the segmentation editor plugin for ImageJ-Fiji[34] (Datasets #1 + #2 and Datasets #9 + #10 in **Supplementary Table 3**). The segmentation labels were chosen to distinguish all relevant diagnoses for the referral decision, as well as potential artefacts that may affect the diagnostic quality of all or part of the scan. In particular, the current state of art does not differentiate between the three different types of pigment epithelial detachment, or segment out areas of fibrosis scarring or blood as hyper-reflective materialdelineations and noclamneture are consistent with standard grading criteria for the evaluation of OCT[27,28]. Anatomical[35–37]. The segmentation examples were selected and segmented by ophthalmologists specializing in medical retina as representative cases for pathological features. These were reviewed and edited by a senior ophthalmologist with over 10 years experience and image reading center certification for OCT segmentation. 3-5 slices per OCT were chosen for segmentation, which best represented the pathological features (**Supplementary Table 2** , **Supplementary Fig. 13, Supplementary Table 9**).

Evaluating the Expert Performance

To evaluate expert performance on the test set, eight clinical experts were recruited for an evaluation study. Participants included four consultant ophthalmologists at Moorfields Eye Hospital with fellowship-level subspecialty training in medical retinal disease and extensive clinical experience (21, 21, 12.5 and 11.5 years of experience respectively), and four optometrists at Moorfields Eye Hospital with specialist training in OCT interpretation and retinal diseases (15, 9, 6 and 2.5 years of experience respectively). These are referred to as retinal specialists 1 to 4 and optometrists 1 to 4 in the rest of the paper (**Supplementary Table 10**) . Each expert was instructed to provide a triage decision (**Supplementary Table 1**) and to record the presence or absence of the defined pathological features (**Supplementary Table 5**).

To assess the performance in a realistic clinical environment, all scans were read in a random order twice with at least a week between readings. On the initial review, only the OCT scan was presented (Dataset #7 in **Supplementary Table 3**). On second review participants were presented with all the information available at the time of triage: OCT and fundus scans, age, gender, ethnicity and where available information on visual acuity and a short clinical vignette (Dataset #8 in **Supplementary Table 3**). The model only received the OCT scan.

To assess the difference between the test set for device type 1 and device type 2 five clinical experts were recruited for a further evaluation study (Dataset #12 in **Supplementary Table 3**). Participants were five consultant ophthalmologists at Moorfields Eye Hospital with fellowship-level subspecialty training in medical retinal disease (21, 21, 12.5, 11.5 and 11 years experience respectively). Four were participants in the device type 1 evaluation study, while the other was a new participant for this study and is referred to as retina specialist five.

Author's version of De Fauw et al. "Clinically applicable deep learning for diagnosis and referral in retinal disease". *Nature Medicine* **XX**, pppp-pppp (2018), DOI: 10.1038/s41591-018-0107-6 **Under Embargo until 13 August 2018 at 1600 London time (DO NOT SHARE)**

Network Architectures and Training Protocol

Segmentation Network

The first stage of our framework consists of a segmentation network that takes as input a part of the OCT scan, and outputs a part of a segmentation map. That is, it predicts for each voxel one tissue type out of the 15 classes described in **Supplementary Table 2**. At training time, the input of the network consists of 9 contiguous slices of an OCT, and the goal of the network is to segment the central slice. The input is therefore a 448x512x9 voxels image, and the output is an estimated probability over the 15 classes, for each of the 448x512x1 output voxels. None of the convolutions made across the slices (z dimension) adds padding to its input. As a result, we can exploit shared computations at inference time to predict any number of contiguous slices in parallel, being only limited by the memory capacity of the system.

The structure of the segmentation convolutional neural network (CNN) model is shown in [14] **Supplementary Fig. 14**. It uses a 3D U-Net architecture , consisting of an analysis (downwards) path, a synthesis (upwards) path, and shortcut connections between blocks of the same level and different paths. We have applied four variations over it. First, we use 3x3x1 convolutions with padding and 1x1x3 convolutions without padding instead of 3x3x3 convolutions without padding. Second, downsampling and upsampling operations are carried out through parameter-free bilinear interpolation, replacing max-pooling and up-convolution. Third, we have introduced one extra residual connection within each block of layers, so that the output of each block consists of the sum of the features of the last layer, and the first layer of the block in which the features dimensions match. Finally, the middle block of layers between the analysis and synthesis paths is composed of a sequence of fully connected layers. The first variation allows us to control the receptive field for z separately and is furthermore less computationally expensive. The second and third variation aimed at improving the gradients flow throughout the network, which makes the training process easier. The last variation extends the receptive field such that each pixel in the output effectively has the whole input contained within its receptive field.

We used per-voxel cross entropy as the loss function, with 0.1 label-smoothing regularization[38]. We have neither used dropout nor weight decay as regularisation means, as preliminary experiments showed this did not improve the performance. We trained the model in TensorFlow[39] with the Adam optimizer[40] for 160000 iterations on 8 Graphics Processing Units (GPUs) with dataset #1 in **Supplementary Table 3** . The initial learning rate was 0.0001 and set to 0.0001/2 after 10% of the total iterations, 0.0001/4 after 20%, 0.0001/8 after 50%, 0.0001/64 after 70%, 0.0001/256 after 90% and finally 0.0001/512 for the final 5% of training. All decisions and hyper-parameters above were selected on the basis of their performance on a validation set (Dataset #2 in **Supplementary Table 3**).

To improve the generalisation abilities of our model we augmented the data by applying affine and elastic transformations jointly over the inputs and ground truth segmentations[13,14]. Intensity transformations over the inputs were also applied.

Our segmentation network for device type 2, shown in **Supplementary Fig. 15** , is trained on scans from both devices (Dataset #1 + #9 in **Supplementary Table 3**) with the aim of leveraging the large number of labeled instances for device type 1. It has three changes compared to the architecture for device type 1. First, we subsample the input from device type 1 (128 slices) to match the resolution of device type 2 (49 slices) and apply slight padding in height to the scans of device type 2 to give them of the same shape in height and width as the scans of device type 1. Second, the input first goes through one of two "device adaptation branches", depending on the device type of the input scan. The architecture of this branch consists of three convolutions with padding, with one residual connection as in the other blocks, and is identical for both device types (see **Supplementary Fig. 15**). The network can then simply learn to compensate for the changes between device types early on and map them to a common representation. Lastly, the number of feature maps on the first level of the analysis path is halved from 32 to 16 such that the overall architecture still has fewer parameters than the architecture for device type 1. During training the network was presented with a ratio of 2.5 : 1 for training samples from device type 2 : device type 1. All decisions and hyper-parameters above were selected on the basis of their performance on a validation set (Dataset #2 + #10 in **Supplementary Table 3**).

Classification network

The classification network learns to map a segmentation map to the four referral decisions and the ten additional diagnoses (see **Supplementary Fig. 16**). For device type 1, it takes as input a 300x350x43 subsampling of the original 448x512x128 segmentation map created by the segmentation network described above. The output is a 14-component vector. For device type 2, whose scans originally are 448x512x49, we first upscale the segmentation map to the same resolution as for device type 1 and then proceed identically as for device type 1. The architecture uses a three-dimensional version of the dense blocks described by Huang et al[41] using 3x3x1 and 1x1x3 convolutions. The details of its structure are shown in **Supplementary Fig. 16**. We found using dense convolution blocks to be critical for training classification networks on large 3d volumes. The inputs are one-hot encoded and augmented by random 3d affine and elastic transformations[14]. The loss was the sum of the softmax cross entropy loss for the first four components (multi-class referral decision) and the sigmoid cross entropy losses for the remaining ten components (additional diagnoses labels). We also used a small amount

(0.05)TensorFlow of label-smoothing regularisation[39403 8]for 160000 iterations of batch size 8 spread across 8 GPUs with 1 and added some (1e-5) weight decay. We trained the model in

with the Adam optimiser

sample per GPU with dataset #3 in **Supplementary Table 3**. The initial learning rate was 0.02 and set to 0.02/2 after 10% of the total iterations, 0.02/4 after 20%, 0.02/8 after 50%, 0.02/64 after 70%, 0.02/256 after 90% and finally 0.02/512 for the final 5% of training. All decisions and hyper-parameters above were selected on the basis of their performance on a validation set (Dataset #4 in **Supplementary Table 3**).

Ensembling

For both of these networks we trained 5 instances: we trained the same network with a different order of the inputs and different random weight initialisations[42]. The experiments of Lakshminarayanan et al[42] suggest that 5 instances are sufficient in most settings, so we also used this number. For our experiments, we applied the 5 instances of our segmentation model to the input scan resulting in 5 segmentation maps. The 5 instances of our classification model were then applied to each of the segmentation maps, resulting in a total of 25 classification outputs per scan, as illustrated in **Supplementary Fig. 1**. The results reported are obtained after averaging the probabilities estimated by these models.

Optimizing the Ensemble Output for Sensitivity, Specificity and Penalty Scores

For different applications, the preferred compromise between a high hit rate (sensitivity) and a low false alarm rate (1-specificity) can be different. For the binary diagnosis decisions, we computed an optimal rescaling factor *a* for the pseudo-probabilities, such that a 50% threshold achieves maximal (sensitivity+specificity)/2 on the validation set (Dataset #4 in **Supplementary Table 3**). The rescaling was done by $p = aq / (aq + (1-a)(1-q))$, where q denotes the ensemble output and p the reweighted probability. We used (sensitivity+specificity)/2 instead of the total accuracy to avoid the bias due to the low number of patients with positive condition in the validation set (and in the test set). For a balanced set with equal numbers of positive and negative samples this term is exactly the accuracy.

For the four-way referral decision (where the highest probability wins), we optimized four scaling factors using the validation set to reduce the overall cost specified by the misclassification penalty matrix (**Supplementary Fig. 6**). A first set of factors was optimized for a balance between high accuracy and low penalty points (referred to as "**our model (1)**" in **Supplementary Fig. 6.**) , a second set of factors was optimized for penalty cost only (referred to as "**our model (2)**" in **Supplementary Fig. 6**). The cost matrix for the balanced performance was computed by averaging the normalized cost matrix for accuracy (a matrix with 0 in the diagonal elements and 1 in the off-diagonal elements) and the normalized penalty cost matrix. Normalisation was performed by dividing the matrix by the sum of all elements. The optimisation of the four factors was done with the Adam optimiser using a softmax layer and a weighted cross-entropy loss layer.

End-to-End Classification Network

The network architecture for the end-to-end classification experiments was identical to the architecture of the classification network in the two-stage approach (see **Section "classification network"** and **Supplementary Fig. 16**) with a small adaption. To roughly obtain the same number of parameters, we added a dense layer (two convolutions with 7 channels output each) that translates the single-channel raw OCT to a 14-channel feature map. All selected hyper parameters and augmentation strategies were identical to the original classification network. We trained 5 network instances on the training set with 14,884 raw OCT scans from device type 1 (Dataset #13 in **Supplementary Table 3**). Each network instance was initialized with different random weights and was presented with the training images in a different order. After training we also computed an optimal reweighting on the validation set (as we did for the two-stage model) and tested the ensemble on the test set.

Statistical Analysis

Significant Differences Using a Two-Sided Exact Binomial Test

The comparison of our model's performance to the expert's performance is based on the assumption that our model and the expert have an unknown but constant performance. That is, every inspected eye scan is correctly diagnosed by our model with the probability p_{mod}, and correctly diagnosed by the expert with probability p_{exp}. For N eye scans the number of correct diagnoses k *is* therefore binomially distributed with $\Pr(k) = \mathrm{B}(k \mid p, N)$. If our model achieves k_{mod} correct diagnoses and the expert achieves k_{exp} correct diagnoses, the probability that the true performance of our model p_{mod} is higher than the true performance of the expert p_{exp} is

$$\Pr(p_{\text{mod}} > p_{\text{exp}} \mid k_{\text{mod}}, k_{\text{exp}}, N) = \frac{\int_0^1 \mathrm{B}(k_{\text{mod}} \mid p_1, N) \cdot \int_0^{p_1} \mathrm{B}(k_{\text{exp}} \mid p_2, N)\, dp_2\, dp_1}{\int_0^1 \mathrm{B}(k_{\text{mod}} \mid p, N)\, dp \int_0^1 \mathrm{B}(k_{\text{exp}} \mid p, N)\, dp}.$$

The probability for a lower performance, i.e. $Pr(p_{\text{mod}} < p_{\text{exp}} \mid k_{\text{mod}}, k_{\text{exp}}, N)$ is derived analogously. For all comparisons, a confidence level of 95% was used. The formula was numerically integrated using in-house code.

Reporting Summary

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

Code Availability

The codebase for the deep learning framework makes use of proprietary components and we are unable to publicly release this code. However, all experiments and implementation details are described in sufficient detail in the methods section and in Supplementary Figures to allow independent replication with non-proprietary libraries. 3D augmentation code (using the caffe framework) is available as part of the 3D U-net source code at https://lmb.informatik.uni-freiburg.de/resources/opensource/unet.en.html. Additionally, although we are unable to make all the Google proprietary components available, we are in the process of making the augmentation operations for TensorFlow available in the official TensorFlow code.

Data availability

The clinical data used for the training, validation and test sets were collected at Moorfields Eye Hospital and transferred to the DeepMind data center in the UK in 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.

Methods-only references

34. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9,** 676–682 (2012).
35. Keane, P. A. *et al* . Evaluation of age-related macular degeneration with optical coherence tomography. *Surv. Ophthalmol.* **57,** 389–414 (2012).
36. Folgar, F. A. *et al.* Comparison of optical coherence tomography assessments in the comparison of age-related macular degeneration treatments trials. *Ophthalmology* **121,** 1956–1965 (2014).
37. Duker, J. S., Waheed, N. K. & Goldman, D. *Handbook of Retinal OCT: Optical Coherence Tomograph y E-Book*. (Elsevier Health Sciences, 2013).
38. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. *P roc IEEE Comput Soc Conf Comput Vis Pattern Recognit* 2818–2826 (2016).
39. Abadi, M. *et al.* TensorFlow: large-scale machine learning on heterogeneous systems. *P reprint at https://arxiv.org/abs/1603.04467* (2016).
40. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. *Preprint at http://arxiv.org/abs/1412.6980* (2014).
41. Huang, G., Liu, Z., Weinberger, K. Q. & van der Maaten, L. Densely connected convolutional networks. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit* **1 ,** 3 (2017).
42. Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Adv. Neural Inf. Process. Syst.* 6405–6416 (2017).
43. De Fauw, J. *et al.* Automated analysis of retinal imaging using machine learning techniques for computer vision. *F1000Res.* **5,** 1573 (2016).

Abbreviations

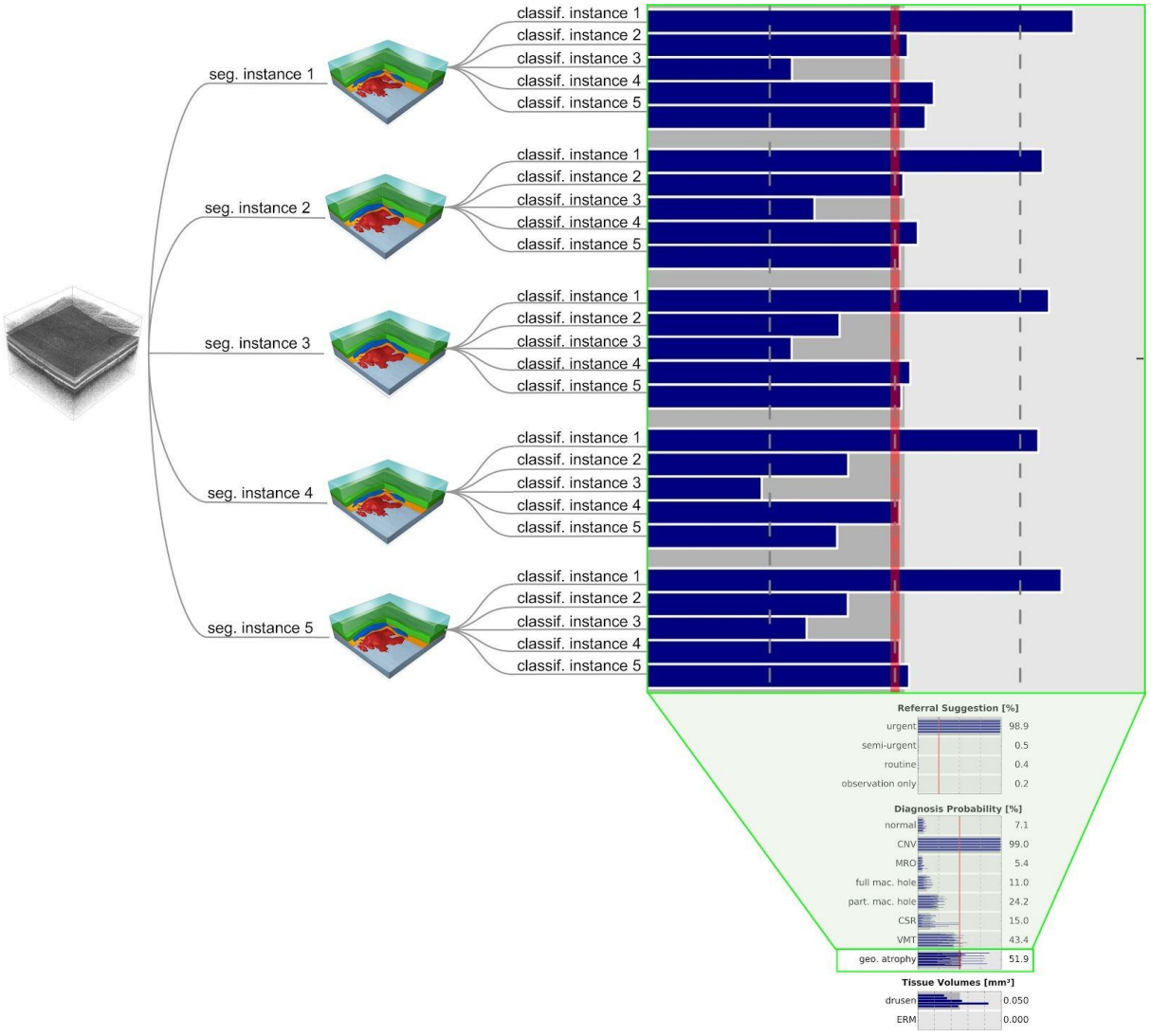
AMD: Age related macular degeneration
AUC: Area under the curve classif.:
classification
CNN: Convolutional neural network
CNV: Choroidal neovascularization
CONSORT: Consolidated Standards of Reporting Trials conv:
convolution
CSR: Central serous retinopathy
DME: Diabetic macular edema
EHR: Electronic health record
ERM: Epiretinal membrane
ETDRS: Early Treatment Diabetic Retinopathy Study
full mac. hole: Full thickness macular hole
GA: Geographic atrophy geo. atrophy:
Geographic atrophy GPU: Graphics
Processing Unit hyper reflect. mat.:
hyperreflective material MacTel: Macular
telangiectasia MEH: Moorfields Eye
Hospital
MRE: Macular retinal edema
NHS: National Health Service

OCT: Optical coherence tomography part.
mac. hole: Partial thickness macular hole
PED: Pigment epithelium detachment perf.:
performance pred.: predicted
QA: Quality Assurance
RAP: Retinal angiomatous proliferation
REC: Research Ethics Committee
ROC: Receiver operating characteristic
RPE: Retinal pigment epithelium
segm.: segmentation Val.: Validation
VMT: Vitreomacular traction
voxel: Volumetric picture element

Author's version of De Fauw et al. "Clinically applicable deep learning for diagnosis and referral in retinal disease". *Nature Medicine* **XX**, pppp-pppp (2018), DOI: 10.1038/s41591-018-0107-6 **Under Embargo until 13 August 2018 at 1600 London time (DO NOT SHARE)**

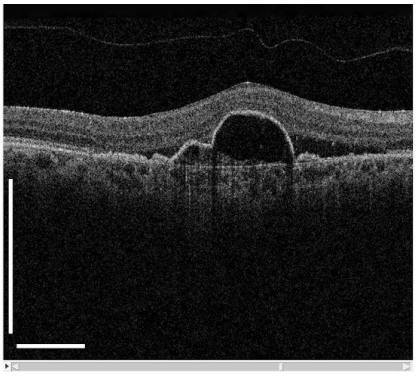
Supplementary Information

Supplementary Figures



Supplementary Figure 1 | Generating predictions with an ensemble of segmentation and classification networks. Illustration showing how the ensemble of 5 segmentation network instances and 5 classification network instances are jointly used to generate 25 predictions for one scan. Each segmentation network instance first provides a segmentation map hypothesis based on the input OCT. For each of these 5 segmentation map hypotheses every classification network instance provides a probability for each label, here shown in detail for the geographic atrophy label.

a

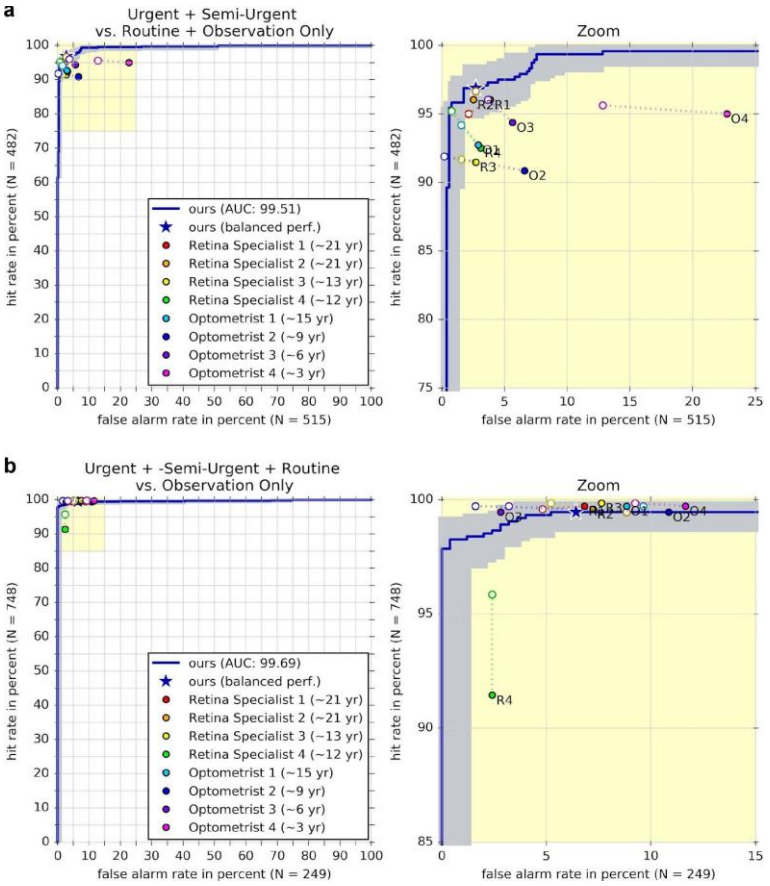


b

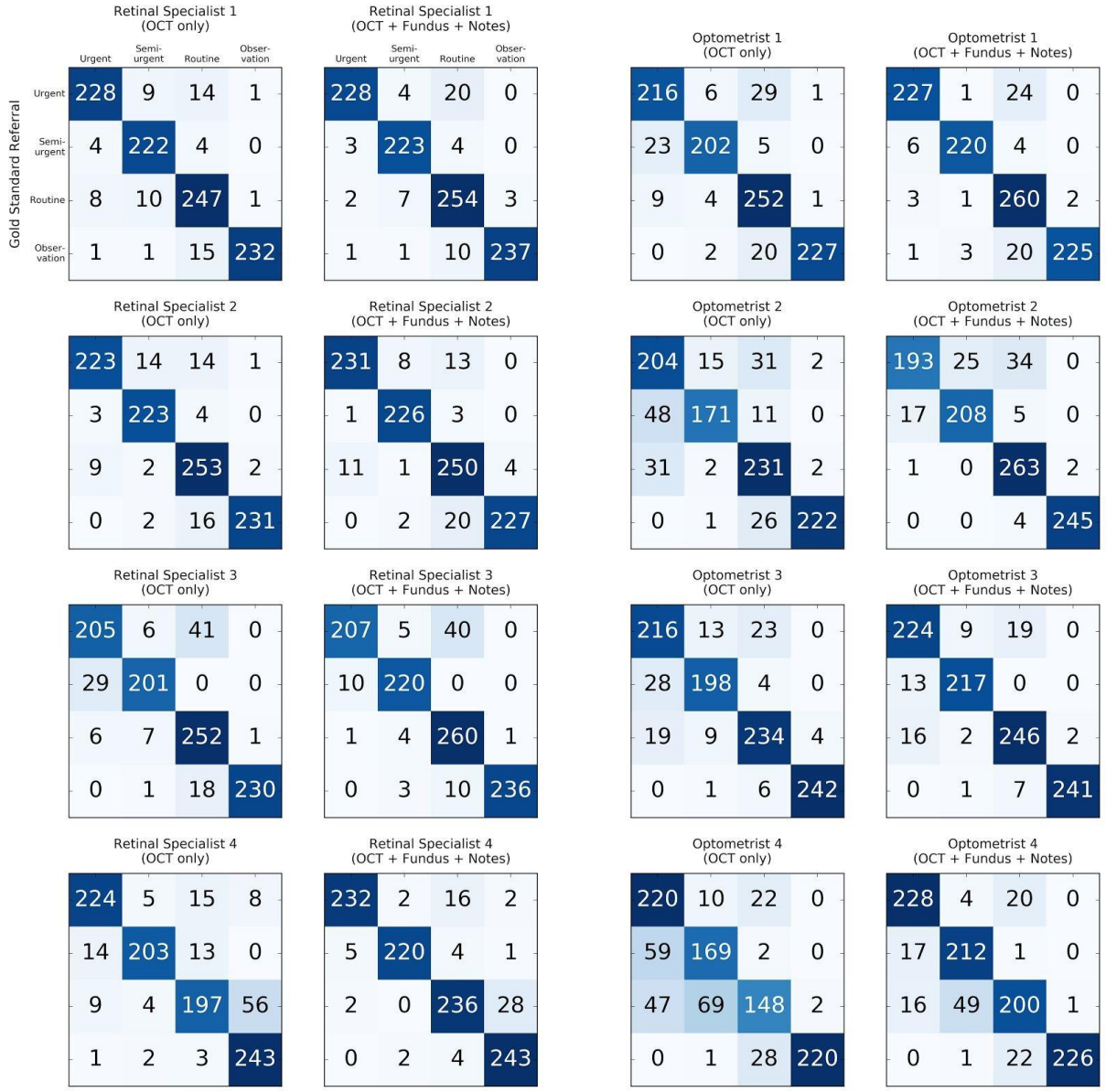
c

Anonymized ID	603425
Age	72
Sex	M
Ethnicity	Asian Indian
Visual Acuity	38
Correction	Glasses
Pinhole	43
Clinical Vignette	Reduced vision

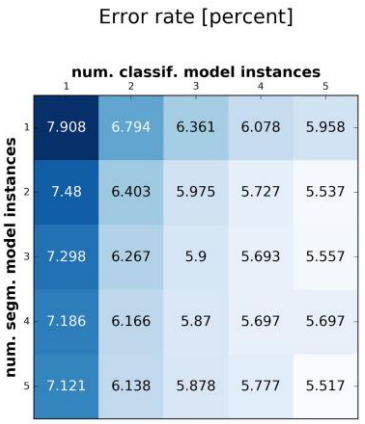
Supplementary Figure 2 | Example information available to the experts to make the referral decision. (**a**) Full OCT scan with a slider to scroll through the slices. (**b**) Fundus image. (**c**) Patient summary notes. Scale bars: 1mm



Supplementary Figure 3 | Receiver operating characteristic (ROC) diagrams for referral decisions. (**a**) Urgent and semi-urgent referral versus routine referral and observation only (n=997 patients). The blue ROC curve is created by sweeping a threshold over the predicted probability (or the measured segmentation volume in case of Drusen and epiretinal membrane (ERM)). Points outside the light blue area correspond to a significantly different performance (95% confidence level, using a two-sided exact binomial test). Filled markers denote expert's performance using OCT only; empty markers denote their performance using OCT, fundus image and summary notes. Dashed lines connect the two performance points of each expert. (**b**) Urgent, semi-urgent and routine referral versus observation only (n=997 patients)

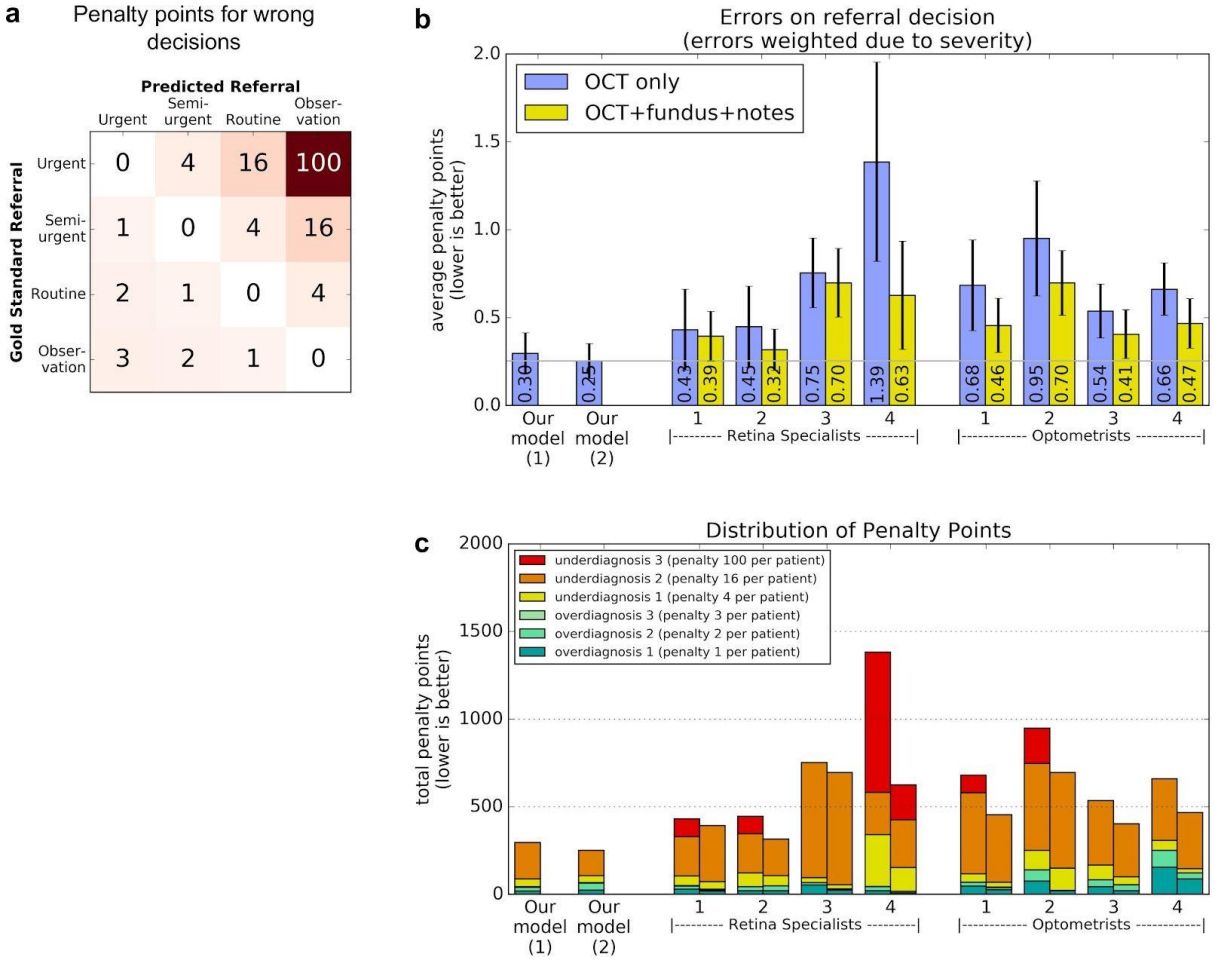


Supplementary Figure 4 | Confusion matrices for the referral decision for all 8 experts. n=997 patients.

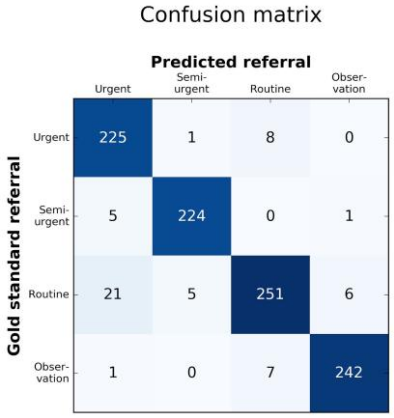


Supplementary Figure 5 | Error rate of the referral decision on device type 1. The error rate is shown for the device type 1 test set (dataset #5 in **Supplementary Table 3**) for different numbers of model instances in the ensemble. The error rate was computed as the average over all possible combinations of N x M segmentation and classification model instances out of the 5 x

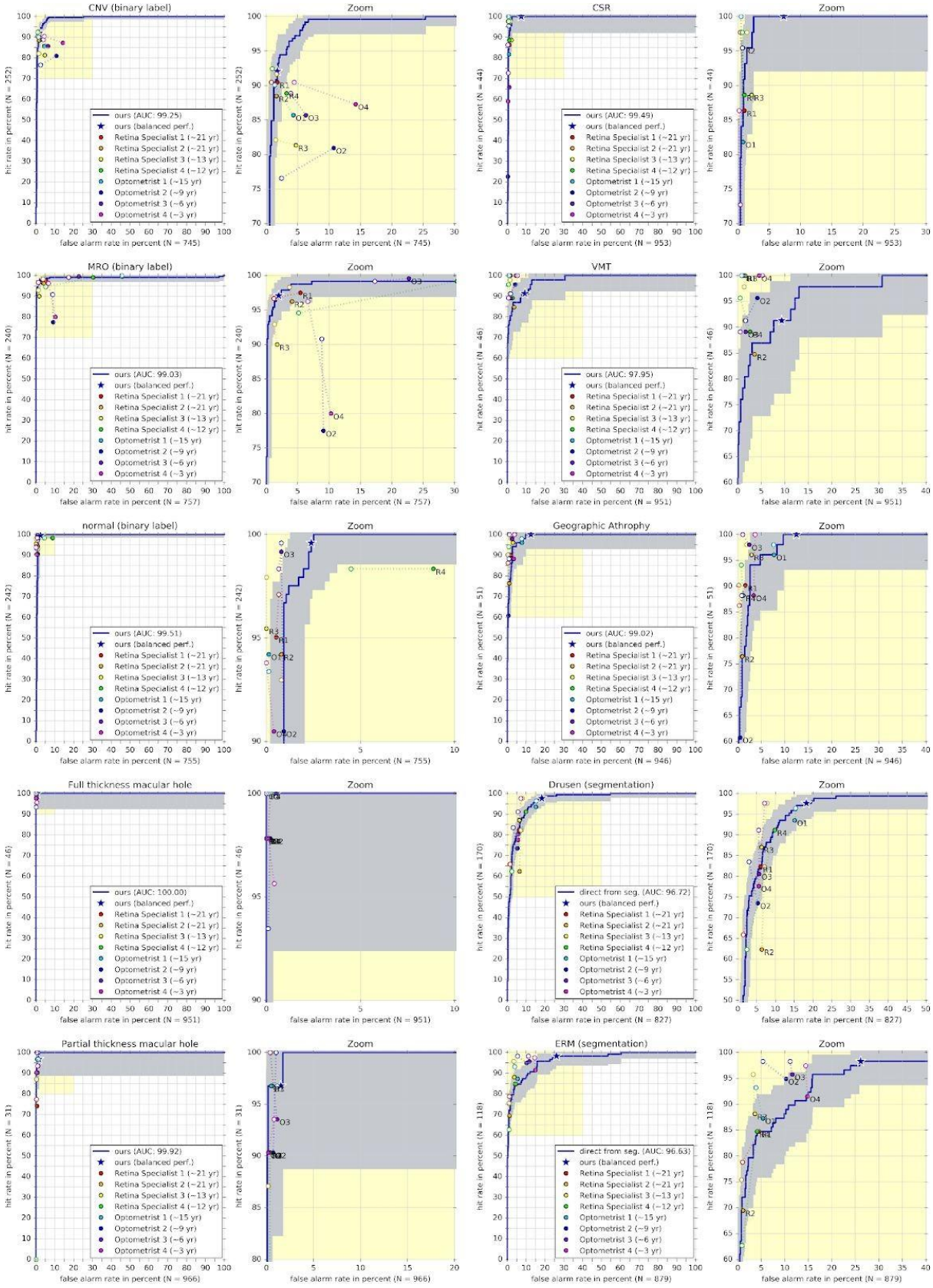
5 instances that are used in the rest of this study. The performance differences between 4 x 4 instances and 5 x 5 instances are only marginal (n=997 patients)



Supplementary Figure 6 | Adverse consequences of wrong referral decisions. (**a**) Our proposed penalty points for a wrong referral decision. In the first row the penalty points correspond approximately to the number of weeks that a CNV patient might lose in referral time before treatment (with 100 as a maximum for a patient triaged as 'observation' and would not be called back for assessment). The penalty points in the other rows are selected relative to this, with the additional constraint that an overdiagnosis (lower left triangle) is considered less harmful to an individual than an underdiagnosis. (**b**) Average penalty points per patient according to our proposed penalty metrics. Framework (1) is optimized for balanced performance; framework (2) is optimized for a better penalty score (n=997 patients, error bars indicate 95% confidence interval, computed from the standard error of the sample mean) (**c**) Distribution of the collected penalty points for our models and the experts (n=997 patients) in the same layout as above. The colored parts of each bar indicate the amount of penalty points collected in each category.

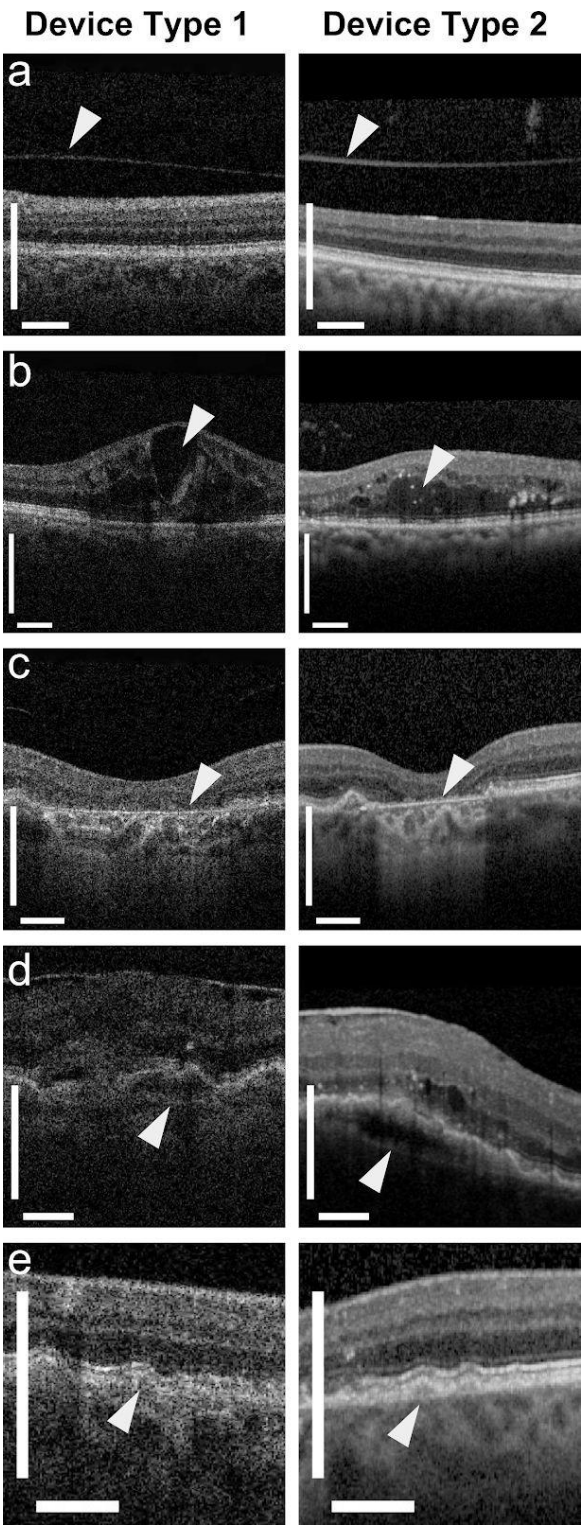


Supplementary Figure 7 | Confusion matrix of an end-to-end classification network applied to the test set (dataset #5 in **Supplementary Table 3**). The results were obtained by an ensemble of 5 model instances (n=997 patients).

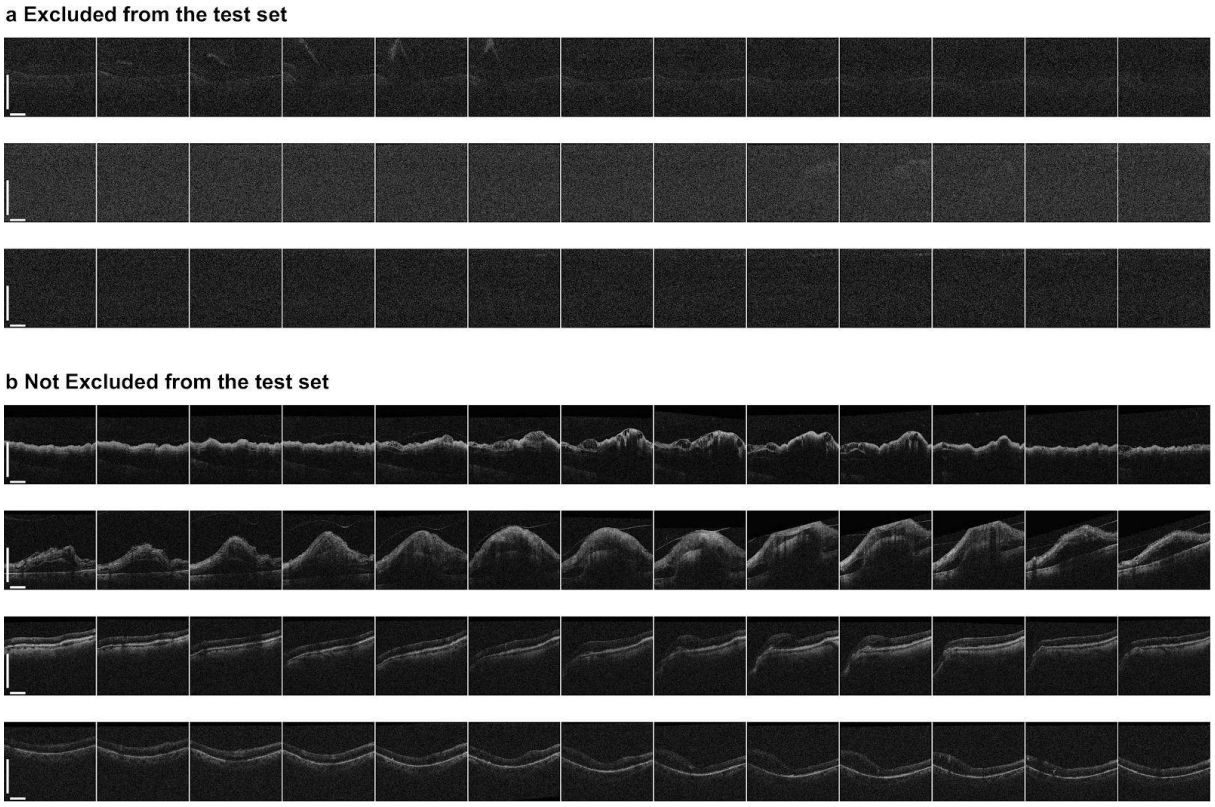


Supplementary Figure 8 | Receiver operating characteristic (ROC) diagrams for the additional pathologies. The blue ROC curve is created by sweeping a threshold over the predicted probability (or the measured segmentation volume in case of Drusen and ERM). n=997 patients. Points outside the light blue area correspond to a significantly different performance (95% confidence level, using a two-sided exact binomial test). Filled markers denote expert's performance using OCT only; empty markers denote their performance using OCT, fundus image and summary notes. Dashed lines connect the two performance points of each expert.

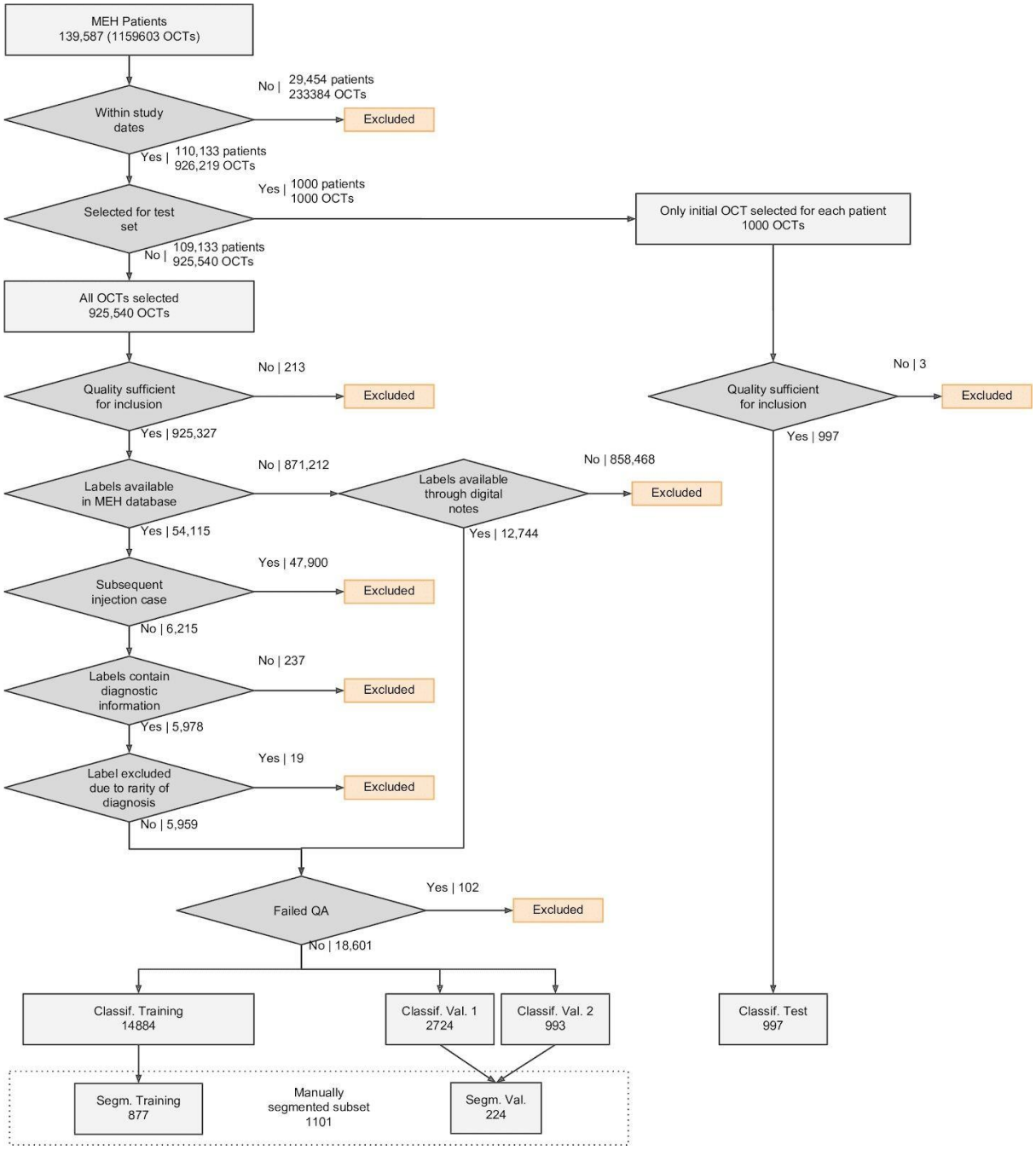
Author's version of De Fauw et al. "Clinically applicable deep learning for diagnosis and referral in retinal disease". *Nature Medicine* **XX**, pppp-pppp (2018), DOI: 10.1038/s41591-018-0107-6 **Under Embargo until 13 August 2018 at 1600 London time (DO NOT SHARE)**



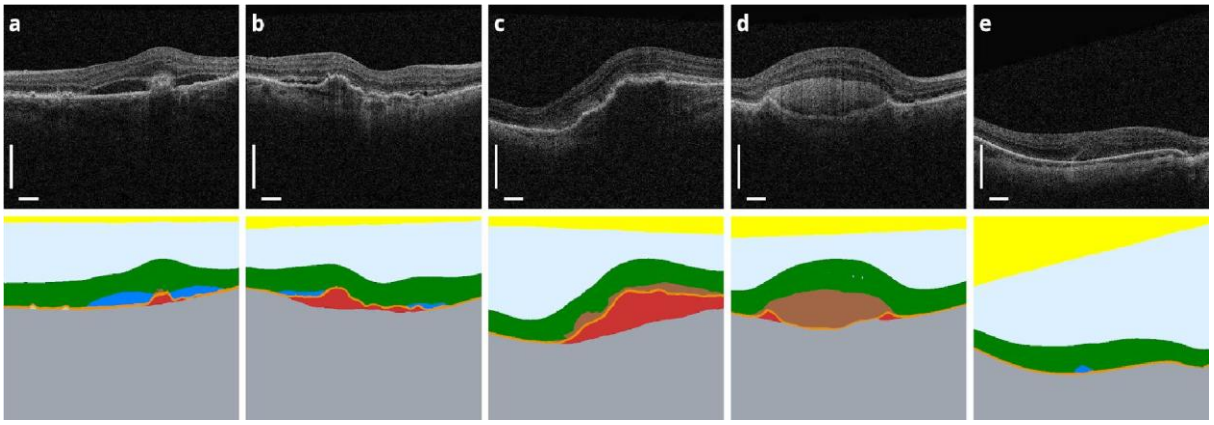
Supplementary Figure 9 | Differences between the two device types. The scans taken by device type 2 (Heidelberg Spectralis) have obvious differences in appearance compared to those taken by device type 1 (Topcon 3D OCT). The higher contrast in device type 2 results in better feature definition which could mislead segmentation models trained only on device type 1. Arrowheads in each image show the differences between device types 1 & 2 respectively. (**a**) Posterior hyaloid. (**b**) Intraretinal fluid. (**c**) Geographic atrophy. (**d**) Fibrovascular PED. (**e**) Drusen. In addition, in all images the choroid - the area below the retina at the bottom of the images - is better defined with device type 2. While this has benefits for diagnosis the differences can confuse models which may mistake the differences for additional retinal layers. Scale bars: 0.5mm



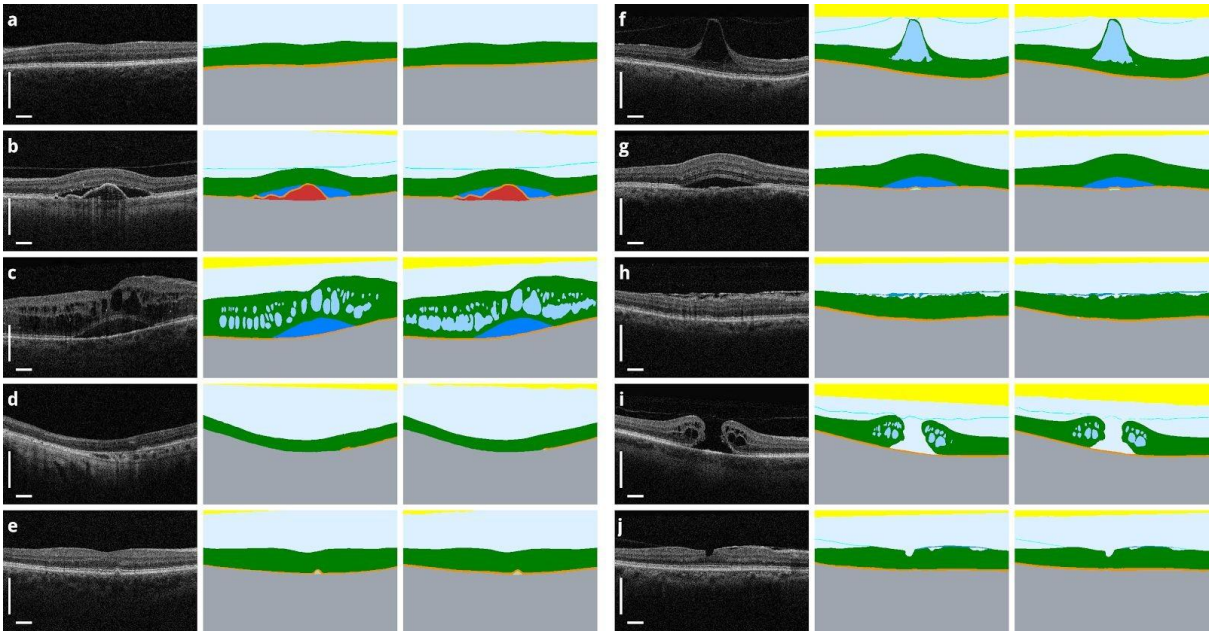
Supplementary Figure 10 | Excluded and included cases in the test set. (**a**) The three cases excluded from the test set due to insufficient signal in the OCT. Every 10th slice of the OCT scan is displayed. Note that in all cases the retina is either absent (empty scan) or barely visible, preventing the interpretation of the scan. (**b**) Four examples of cases that were included in the test set despite being of poor quality, or representing complex pathology. Scale bars: 1mm



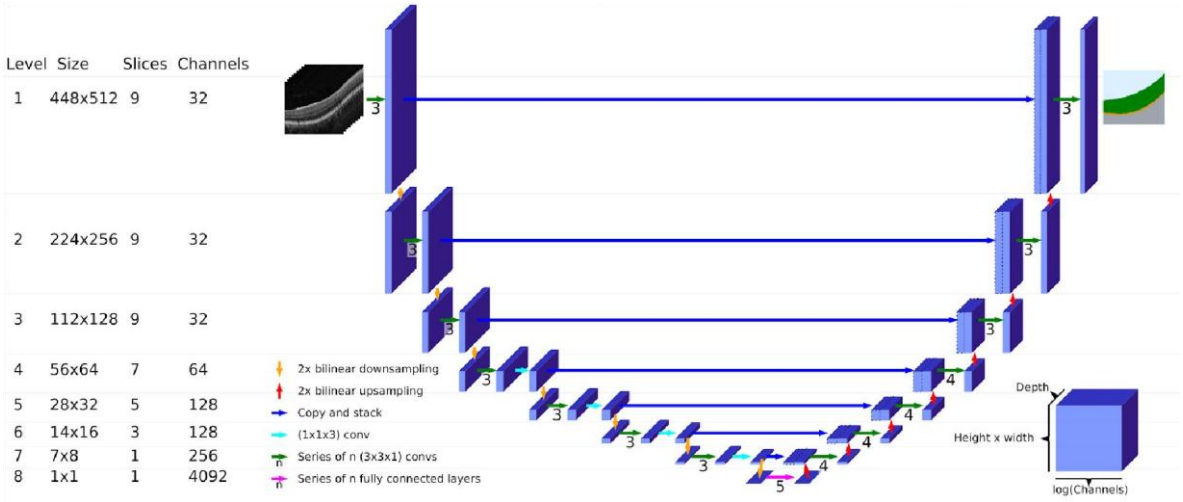
Supplementary Figure 11 | Sample selection at Moorfields Eye Hospital (MEH). Manual opt outs are not included as none of the patients who manually opted out had digital OCT within the study dates.



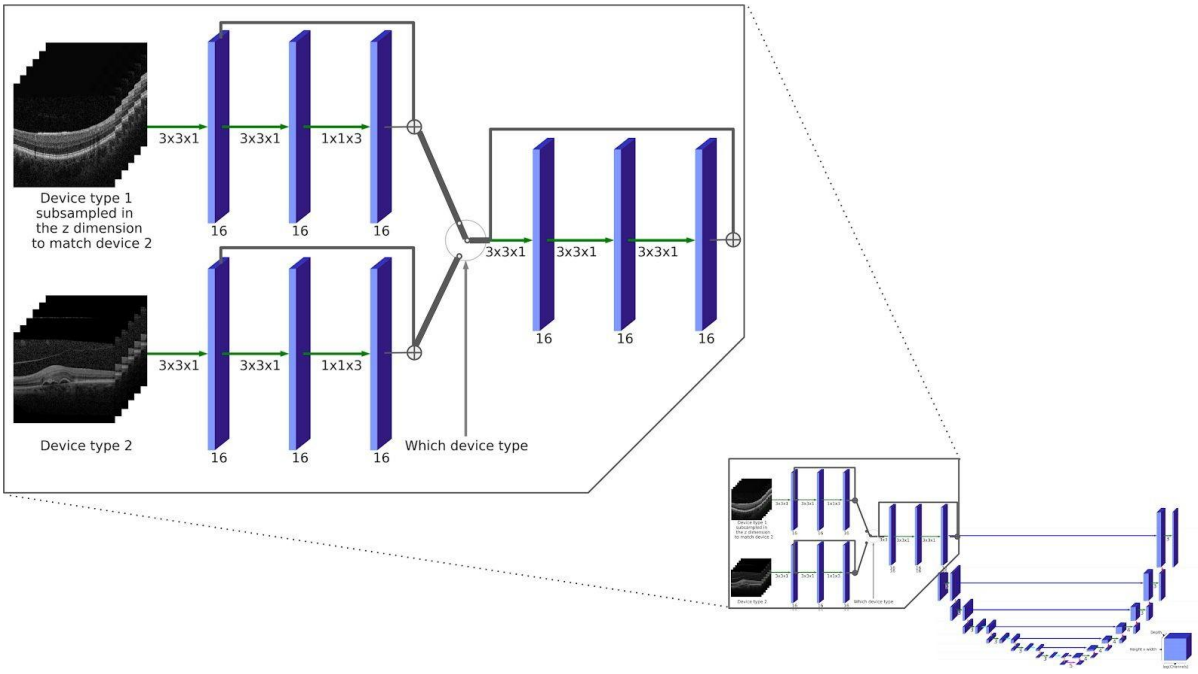
Supplementary Figure 12 | Five examples of patients in the test set with choroidal neovascularization (CNV). All images from the test set (n=997) show CNV requiring urgent referral with corresponding segmentations from our segmentation network (color legend in **Supplementary Table 2)**. (**a**) A patient with choroidal neovascularization in the context of CSR. (**b**) A patient with choroidal neovascularization resulting from age related macular degeneration (AMD). **(c)** A patient with extensive fibrovascular pigment epithelium and subretinal hyperreflective material. (**d**) A patient with large amounts of subretinal hyperreflective material in the context of CNV. (**e**) A highly ambiguous case with a possible retinal angiomatous proliferation (RAP) lesion in a myopic patient. Scale bars: 0.5mm



Supplementary Figure 13 | Examples of the segmentation model output for ten different retinal pathologies. 2D slices of OCT scans in the segmentation test set (n=224) with corresponding manual and predicted segmentation maps for the ten pathology classes included in our study (color legend in **Supplementary Table 2)**. Classes are not mutually exclusive and multiple pathologies may be present in a single scan. (**a**) A normal retina as it appears in an OCT scan. (**b**) A patient with choroidal neovascularization due to age related macular degeneration. The segmentation map shows the area of fibrovascular pigment epithelium detachment associated with neovascularization. (**c**) Diabetic maculopathy and referable macular edema. (**d**) Geographic atrophy in late age related macular degeneration. (**e**) Drusen in early age related macular degeneration. (**f**) Vitreomacular traction. (**g**) Central serous retinopathy. Note that the segmentation model correctly identifies the pigment epithelium detachment as serous material. (**h**) A patient with epiretinal membrane. (**i**) Full thickness macular hole. (**j**) Partial thickness macular hole. Scale bars: 0.5mm

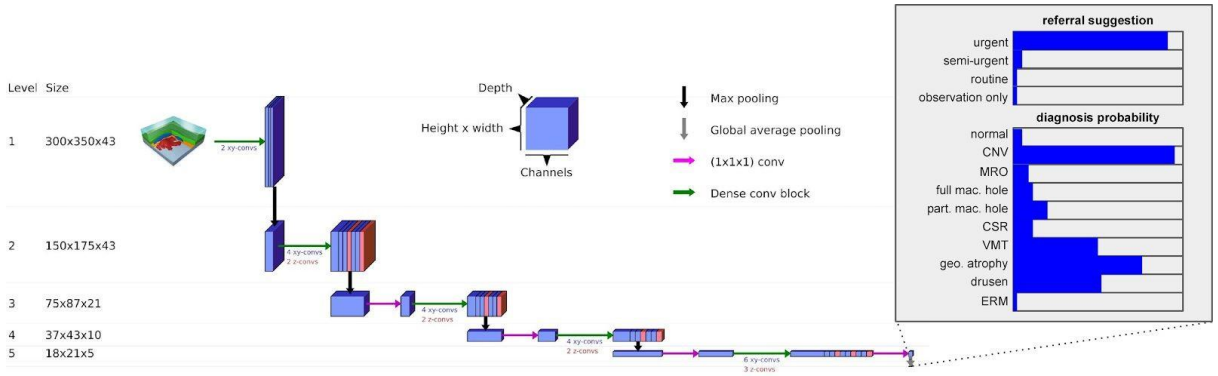


Supplementary Figure 14 | 3D U-Net model used in the first stage of our approach. At training time, the model receives 9 contiguous OCT slices. Blue boxes illustrate the 4D activation maps. Colored arrows stand for the different operations.



Supplementary Figure 15 | 2-branch U-Net for device type 2. The architecture of our segmentation network with "device adaptation branches" to segment scans of device type 2. In the top left we show an enlarged version of the differences compared to the original architecture for device type 1 (shown in **Supplementary Fig. 14**). Blue boxes illustrate the 4D activation maps with the number of channels shown below. Green arrows denote convolutional operations. We train on scans from both device type 1 and device type 2 but subsample those from device type 1 in the z-dimension to match the lower z-resolution of device type 2. Depending on which device the scan is from, the scan first goes through either the top branch, for device type 1, or the bottom branch, for device type 2. The output of the chosen branch is then used as input to a modified version of the first level of the analysis path of the original architecture. The rest of the architecture is identical.

Author's version of De Fauw et al. "Clinically applicable deep learning for diagnosis and referral in retinal disease". *Nature Medicine* **XX**, pppp-pppp (2018), DOI: 10.1038/s41591-018-0107-6 **Under Embargo until 13 August 2018 at 1600 London time (DO NOT SHARE)**



Supplementary Figure 16 | Classification CNN (convolutional neural network) used in the second stage of our approach . Blue and red boxes illustrate the 4D activation maps. Blue boxes are the result of a (3x3x1) convolution, while red boxes are the result of a (1x1x3) convolution.

Author's version of De Fauw et al. "Clinically applicable deep learning for diagnosis and referral in retinal disease". *Nature Medicine* **XX**, pppp-pppp (2018), DOI: 10.1038/s41591-018-0107-6 **Under Embargo until 13 August 2018 at 1600 London time (DO NOT SHARE)**

Supplementary Tables

Supplementary Table 1 | Taxonomy of referral classes

Referral Category	Definition
Urgent	All causes of choroidal neovascularization, including age related macular degeneration, high myopia, central serous retinopathy, inherited retinal dystrophies (e.g., angioid streaks), posterior uveitis (e.g., multiple choroiditis), and post traumatic choroidal rupture.
Semi-urgent	Referable edema classed as semi-urgent included diabetic maculopathy, retinal vein occlusion, postoperative (Irvine-Gass syndrome), uveitis, Coat's disease, radiation and miscellaneous other cases.
Routine	All other non-urgent cases with a large variety, from uncomplicated central serous retinopathy to more rare conditions such as Macular Telangiectasia (MacTel) type 2.
Observation only	The absence of pathology classes described above.

Author's version of De Fauw et al. "Clinically applicable deep learning for diagnosis and referral in retinal disease". *Nature Medicine* **XX**, pppp-pppp (2018), DOI: 10.1038/s41591-018-0107-6 **Under Embargo until 13 August 2018 at 1600 London time (DO NOT SHARE)**

Supplementary Table 2 | Taxonomy of segmentation regions

			Training		Test	
Color	**Feature**	**Definition**	**Total number of scans with label segmented**	**Percent of total voxels**	**Total number of scans with label segmented**	**Percent of total voxels**
	Vitreous and subhyaloid	Area above the internal limiting membrane not covered by other segmentation classes	856	20.63	220	20.30
	Posterior hyaloid	Hyperreflective membrane visible above the retina in cases of posterior vitreous detachment	356	0.12	95	0.13
	Epiretinal membrane	Hyperreflective band seen on the inner surface of the retina, often associated with distortion of the underlying neurosensory retina	326	0.06	95	0.04
	Neurosensory retina	All layers and contents of the retina, excluding the pathological features described below	856	10.76	220	11.12
	Intraretinal fluid	Areas of round or oval hyporeflectivity located within the neurosensory retina	356	0.59	111	0.76
	Subretinal fluid	Hyporeflective areas in the subretinal space - the space below the neurosensory retina but above the retinal pigment epithelium	255	0.48	68	0.33
	Subretinal hyper reflective material	Areas of hyperreflectivity between the retinal and RPE	92	0.12	22	0.08
	Retinal pigment epithelium (RPE)	Hyperreflective band underlying the neurosensory retina	853	0.92	220	0.95
	Drusenoid pigment epithelium detachment (PED)	Elevation of the RPE, often dome-shaped, with a hypo- or medium-reflective material separating the RPE from the underlying Bruch's membrane, and without the presence of fibrovascular material	268	0.07	61	0.06
	Serous PED	Dome-shaped elevation of the retinal pigment epithelium relative to Bruch's membrane, typically seen overlying a homogeneously hyporeflective space devoid of fibrovascular material	58	0.02	11	0.003
	Fibrovascular PED	Irregular elevations of the retinal pigment epithelium relative to Bruch's membrane containing fibrovascular tissue of variable reflectivity	183	0.37	45	0.54
	Choroid and outer layers	Area below the RPE not covered by other segmentation classes	854	51.26	220	51.56

	Mirror artefact	Artefact caused by patient anatomy out of the OCT frame being reflected back onto the OCT	9	0.01	3	0.02
	Clipping artefact	Padding voxels introduced at the edges of OCT slices during image processing	877	7.21	224	7.36
	Blink artefact	Absent information due to patient blink	22	7.38	5	6.75

Supplementary Table 3 | Overview of datasets used for training, validation and testing of the different networks

	Dataset	Device type	Number of scans	Input	Labels	Label source
#1	Training set for segmentation	1	877	OCT scans	Sparse segm. maps (3-5 slices per scan)	Manually segmented by trained ophthalmologists, reviewed and edited by a senior ophthalmologist.
#2	Validation set for segmentation	1	224	OCT scans	Sparse segm. maps (3-5 slices per scan)	Manually segmented by trained ophthalmologists, reviewed and edited by a senior ophthalmologist.
#3	Training set for classification	1	14884	dense segm. maps, created automatically from OCT scans (5 segm. maps per scan)	Diagnoses and referral decision	Automated notes search + trained ophthalmologist and optometrist review of the OCT scans.
#4	Validation set for classification	1	993	dense segm. maps, created automatically from OCT scans	Diagnoses and referral decision	Graded by three junior graders. Disagreement in clinical labels arbitrated by a senior grader.
#5	Test set: Referral gold standard	1	997	OCT scans	Referral decision	Full patient clinical records to determine the final diagnosis and optimal referral pathway in the light of that (subsequently obtained) information.
#6	Test set: Diagnoses silver standard		(same OCT scans as #5)		Diagnoses	Majority vote from 8 experts (4 retinal specialists and 4 optometrists) grading using OCT scan, fundus image and clinical notes
#7	Human results: Experts on OCT only		(same OCT scans as #5)		Diagnoses and referral decision from 8 experts	8 experts (4 retinal specialists and 4 optometrists) grading on OCT scan only
#8	Human results: Experts on OCT + fundus + notes		(same OCT scans as #5)		Diagnoses and referral decision from 8 experts	8 experts (4 retinal specialists and 4 optometrists) grading on OCT scan, fundus image and clinical notes

#9	Training set 2 for segmentation	2	152	OCT scans	Sparse segm. maps (3-5 slices per scan)	Manually segmented by trained ophthalmologists, reviewed and edited by a senior ophthalmologist.
#10	Validation set 2 for classification	2	112	OCT scans	Referral decision	Full patient clinical records to determine the final diagnosis and optimal referral pathway in the light of that (subsequently obtained) information.
#11	Test set 2: Referral gold standard	2	116	OCT scans	Referral decision	Full patient clinical records to determine the final diagnosis and optimal referral pathway in the light of that (subsequently obtained) information.
#12	Human results: Experts on OCT + fundus + notes	(same OCT scans as #11)			Referral decision from 5 experts	5 retinal specialists grading on OCT scan, fundus image and clinical notes
#13	Training set for end-to-end model	(same cases as #3)		OCT scans	(same labels as #3)	

Supplementary Table 4 | Overview of the OCT scan sizes used in this study. All 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
device type 1 raw OCT scans	885 · 512 · 128	2.6 · 11.7 · 47.2	2.3 · 6.0 · 6.0	
segmentation network input / output	448 · 512 · 128	5.2 · 11.7 · 47.2		device type 1 scans resampled in A-scan direction to 5.2µm voxel size, and zero-padded to the next multiple of 64 (added 6 pixels)
classification network 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
device type 2 raw OCT scans	496 · 512 · 49	3.9 · 11.3 · 120	1.93 · 5.79 · 5.88	
2-branch segmentation network input / output	448 · 512 · 49	5.2 · 11.7 · 120		device type 2 scans resampled in A,B-scan direction to 5.2µm · 11.7µm voxel size, and padded accordingly; device type 1 scans resampled in C-scan direction to 120µm voxel size.

Author's version of De Fauw et al. "Clinically applicable deep learning for diagnosis and referral in retinal disease". *Nature Medicine* **XX**, pppp-pppp (2018), DOI: 10.1038/s41591-018-0107-6 **Under Embargo until 13 August 2018 at 1600 London time (DO NOT SHARE)**

Supplementary Table 5 | Taxonomy of diagnostic labels

Condition	Definition
Normal	Absence of pathology.
Macular retinal edema (MRE)	Referable retinal edema, seen in the OCTs as intraretinal and subretinal fluid.
Choroidal neovascularization (CNV)	New vessel growth from the choroidal layer of the eye; associated with a variety of retinal conditions including neovascular age related macular degeneration, severe myopia and central serous retinopathy.
Drusen	Acellular polymorphous deposits in Bruch's membrane; the most common early sign of dry age-related macular degeneration.
Geographic atrophy	Loss of the retinal pigment epithelium with variable loss of the overlying photoreceptors and underlying choriocapillaris; a sign of late stage dry age-related macular degeneration.
Central serous retinopathy (CSR)	A disease where increased choroidal permeability leads to a build up of subretinal fluid, causing a detachment of the neurosensory retina.
Full thickness macular hole	A round, full-thickness defect of retinal tissue in the foveal retina, leading to loss of central vision.
Partial thickness macular hole	A partial thickness defect of retinal tissue in the foveal retina.
Vitreomacular traction (VMT)	A disorder of the vitreoretinal interface where an incomplete posterior vitreous detachment exerts tractional pull on the macula and results in morphologic alterations and consequent metamorphopsia or central visual loss.
Epiretinal membrane (ERM)	Fibrocellular tissue found on the inner surface of the retina which may be idiopathic or secondary to various retinal conditions. Small epiretinal membranes may not be clinically significant, and may be considered a normal aging feature.

Supplementary Table 6 | Same as table before but experts have access to OCT + fundus image + full summary notes.

Diagnosis	Area under ROC curve [percent]	N positive samples	Experts with significantly higher performance	Experts with indistinguishable performance	Experts with significantly lower performance
CNV	99.25	252	-	●●●▲▲	●▲▲
MRE	99.03	240	-	●●●▲▲	●▲▲
normal	99.51	242	●●▲▲▲	●●	▲
full mac. hole	100.0	46	-	●●●●▲▲▲▲	-
part.mac. hole	99.92	31	-	●●●●▲▲▲▲	-
CSR	99.49	44	●●●▲▲	●▲▲	-
VMT	97.95	46	●●●▲▲▲	●▲	-
geographic atrophy	99.02	51	●●●▲▲▲	●▲	-
Drusen	97.42 (from segm.)	170	●●▲▲	●●▲▲	-
ERM	96.63 (from segm.)	118	●●▲▲▲	●●▲	-

Supplementary Table 7 | Performance on additional diagnoses of experts using OCT only to our framework. The ground truth for CNV is derived from the full follow-up patient files. The ground truth for the other diagnoses is computed as majority vote from the 8 experts. Drusen and ERM predictions were derived directly from the segmentation map. Circles represent retinal specialists, triangles represent optometrists.

Diagnosis	Area under ROC curve [percent]	N positive samples	Experts with significantly higher performance	Experts with indistinguishable performance	Experts with significantly lower performance
CNV	99.25	252	-	●●	●●▲▲▲▲
MRE	99.03	240	-	●●▲▲▲	●●▲
normal	99.51	242	●▲▲	●●●▲	▲
full mac. hole	100.0	46	-	●●●●▲▲▲▲	-
part. mac. hole	99.92	31	-	●●▲▲▲▲	●
CSR	99.49	44	-	●●●●▲▲▲▲	-
VMT	97.95	46	●●●▲	●▲▲▲	-
geographic atrophy	99.02	51	▲	●●●●▲▲▲	-
Drusen	97.42 (from segm.)	170	-	●●▲▲▲▲	●
ERM	96.63 (from segm.)	118		●●●▲▲▲▲	-

Supplementary Table 8 | Total OCT examinations (unique patients in brackets) in the dataset by triage category.

Triage category	Training	Validation	Test (Device Type 1)	Test (Device Type 2)

Urgent	4832 (3039)	251 (237)	252 (252)	34 (34)
Semi-urgent	3438 (1854)	268 (259)	230 (230)	28 (28)
Routine	5223 (1927)	247 (236)	266 (266)	35 (35)
Observation only	1391 (801)	227 (195)	249 (249)	19 (19)
Total	**14884 (7621)**	**993 (927)**	**997 (997)**	**116 (116)**

Supplementary Table 9 | Number of cases of referral classes in the training and validation set for the segmentation network for OCT device type 1.

Referral Category	Number in Training Set	Number in Validation Set
Urgent	227	58
Semi-Urgent	182	57
Routine	89	20
Observation	379	89
Total	**877**	**224**

Supplementary Table 10 | The experience and position of the nine experts against which the algorithm was compared.

Expert	Position	Years of Experience
1	Consultant Ophthalmologist in Medical Retina	21
2	Consultant Ophthalmologist in Medical Retina	21
3	Consultant Ophthalmologist in Medical Retina	12.5
4	Consultant Ophthalmologist in Medical Retina	11.5
5	Specialist Optometrist, Medical Retina	15
6	Specialist Optometrist, Medical Retina	9

7	Specialist Optometrist, Medical Retina	6
8	Specialist Optometrist, Medical Retina	2.5
9	Consultant Ophthalmologist in Medical Retina	10

Author's version of De Fauw et al. "Clinically applicable deep learning for diagnosis and referral in retinal disease". *Nature Medicine* **XX**, pppp-pppp (2018), DOI: 10.1038/s41591-018-0107-6 **Under Embargo until 13 August 2018 at 1600 London time (DO NOT SHARE)**

Supplementary Videos

Supplementary Video 1 - OCT viewer | This video demonstrates the interaction with the OCT viewer. The OCT scan belongs to a 72 year old female presented with increasing visual distortion over a 4 month period; the OCT shows loss of RPE consistent with geographic atrophy. The view first goes through the whole volume (128 slices) for a fixed tissue map hypothesis, followed by showing the different tissue map hypotheses for a given slice. Finally, we let the collage cycle through the different hypotheses continually while scrolling through the volume, pausing on several slices briefly to show the variations. The color legend for all segmentation maps is available in **Supplementary Table 2.**

Supplementary Video 2 - wet AMD | Choroidal neovascularization (CNV) is the pathognomonic feature of the neovascular ("wet") form of age-related macular degeneration (AMD) and requires urgent treatment to prevent irreversible visual loss. A 72-year old man presented with a history of reduced vision in his left eye. Best corrected visual acuity in the affected eye was 38 Early Treatment Diabetic Retinopathy Study (ETDRS) letters. The model correctly selects the Most Urgent Diagnosis as "CNV", suggesting referral to an ophthalmologist on an urgent basis. The model segmentation highlights growth of the neovascular tissue in the sub-retinal pigment epithelium (RPE) space – a so-called fibrovascular pigment epithelium detachment (PED). Subretinal fluid can be seen surrounding the inferior margins of the fibrovascular PED indicating the presence of ongoing CNV leakage.

Supplementary Video 3 - Normal | Scans are quick and safe to perform and are thus commonly used in the screening of patients without visual symptoms or other ophthalmic findings. A 46-year old man who was referred for retinal specialist review. Best corrected visual acuity was 6/6. The model correctly selects the referral decision as "Observation Only", suggesting that the OCT findings in isolation do not require referral to an ophthalmologist. The model accurately delineates the neurosensory retina without the presence of any pathologic compartments. It also highlights partial separation of the posterior hyaloid of the vitreous – this is a normal finding as the vitreous gel increasingly liquefies with age.

Supplementary Video 4 - Diabetic Macular Edema | Accumulation of this fluid in the macula – diabetic macular edema (DME) – is the commonest cause of visual impairment in diabetes. A 54-year old man with diabetes was referred to Moorfields for ophthalmologist review with best corrected visual acuity in the affected eye of 45 ETDRS letters. The model correctly detects the presence of macular retinal edema (MRE) and suggests semi-urgent ophthalmology referral. The model highlights intraretinal fluid accumulation, with cystoid spaces in both the inner nuclear and outer plexiform layers, and a mixed petaloid/honeycomb appearance on the en face images. There is also an accompanying significant increase in total retinal thickness.

Supplementary Video 5 - Ambiguous Case (chronic CSR) | In chronic CSR, diagnosis of secondary CNV formation is often challenging due to the frequent presence of shallow irregular pigment epithelium detachments (PEDs). A 60-64 year old woman presented with a history of CSR in her left eye. The model correctly detects the presence of CSR but is far less certain about the presence of CNV. It highlights a gravitational tract of subretinal fluid with a discrete area of fibrovascular PED superior to the fovea.

Supplementary Video 6 - Ambiguous Case (advanced geographic atrophy) | In advanced forms of AMD, geographic atrophy (GA) may sometimes coexist with CNV formation. In such cases, the CNV component may be clinically silent, and the fundus appearance may be limited to that of GA, making the diagnosis difficult. A 84-year old man was referred to Moorfields. Best corrected visual acuity in the affected eye was 1/60. The ground truth diagnosis was GA and routine referral was recommended. While the model correctly diagnoses the presence of GA and drusen, it suggests urgent referral due to the possible presence of CNV. The presence of subretinal hyperreflective on model segmentation is suggestive of previous CNV formation.

Supplementary Video 7 - Difficult Case of CNV | A 30 year old male patient, with a known history of CSR, presented with acute visual loss in his left eye and was diagnosed with secondary CNV formation. At this visit, the OCT scans lack many of the prototypical features of CSR, such as subretinal fluid accumulation. The model correctly diagnoses the presence of CNV and suggests the presence of CSR, but with far less certainty.

Supplementary Video 8 - Failure case (partial thickness macular hole) | Ocular media opacities may sometimes cause artefactual reductions in OCT signal strength and this can make accurate image segmentation challenging. Due to localized reduction in OCT signal strength in this case, some of the models erroneously detect the presence of a partial thickness macular hole. As a result, the models are uncertain as to whether the eye is normal or whether routine referral is required.

Supplementary Video 9 - Integration with other clinical information | Retinal angiomatous proliferation (RAP) is a variant of choroidal neovascularization (CNV) due to age-related macular degeneration (AMD). A 75-79 year old woman presented with reduced vision in her left eye. The model segmentation highlights the presence of a fibrovascular pigment epithelium detachment (PED) with subretinal hyperreflective material, overlying intraretinal fluid, and surrounding drusen. These findings

are highly suggestive of RAP - in its early stages, this can be misdiagnosed as macular retinal edema (MRE), particularly in elderly patients with diabetes. The interpretable representation reduces the risk of misdiagnosis and allows the clinician to easily correlate these findings with other clinical information, e.g., fundus fluorescein angiography.

Author information

Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to O.R. (olafr@deepmind.com), P.A.K. (pearse.keane@moorfields.nhs.uk)