


Automated artificial intelligence-based system for clinical follow-up of patients with age-related macular degeneration

Ivan Potapenko,^{1,2}  Bo Thiesson,^{3,4} Mads Kristensen,³ Javad Nouri Hajari,¹ Tomas Ilginis,¹ Josefine Fuchs,¹ Steffen Hamann^{1,2} and Morten la Cour^{1,2}

¹Department of Ophthalmology, Rigshospitalet, Copenhagen, Denmark

²Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

³Enversion A/S, Aarhus, Denmark

⁴Department of Engineering, Aarhus University, Aarhus, Denmark

ABSTRACT.

Purpose: In this study, we investigate the potential of a novel artificial intelligence-based system for autonomous follow-up of patients treated for neovascular age-related macular degeneration (AMD).

Methods: A temporal deep learning model was trained on a data set of 84 489 optical coherence tomography scans from AMD patients to recognize disease activity, and its performance was compared with a published non-temporal model trained on the same data (Acta Ophthalmol, 2021). An autonomous follow-up system was created by augmenting the AI model with deterministic logic to suggest treatment according to the observe-and-plan regimen. To validate the AI-based system, a data set comprising clinical decisions and imaging data from 200 follow-up consultations was collected prospectively. In each case, both the autonomous AI decision and original clinical decision were compared with an expert panel consensus.

Results: The temporal AI model proved superior at detecting disease activity compared with the model without temporal input (area under the curve 0.900 (95% CI 0.894–0.906) and 0.857 (95% CI 0.846–0.867) respectively). The AI-based follow-up system could make an autonomous decision in 73% of the cases, 91.8% of which were in agreement with expert consensus. This was on par with the 87.7% agreement rate between decisions made in the clinic and expert consensus ($p = 0.33$).

Conclusions: The proposed autonomous follow-up system was shown to be safe and compliant with expert consensus on par with clinical practice. The system could in the future ease the pressure on public ophthalmology services from an increasing number of AMD patients.

Key words: artificial intelligence – age-related macular degeneration – anti-vegf – follow-up

Acta Ophthalmol. 2022; 100: 927–936

© 2022 The Authors. Acta Ophthalmologica published by John Wiley & Sons Ltd on behalf of Acta Ophthalmologica Scandinavica Foundation.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

doi: 10.1111/aos.15133

Introduction

Current treatment strategies for neovascular macular degeneration (AMD) markedly improve visual outcomes (Brynskov *et al.* 2020, Papadopoulos 2020), but often require close monitoring and administration of anti-VEGF injections for extended periods of time (Baek *et al.* 2019). Together with the demographic transition to an older population, this will continue to be the main driving force behind the rise in the number of patients in need of treatment (Potapenko & la Cour 2021), increasing the burden on ophthalmology services. If the quality of care is to be maintained in the future, novel approaches are needed for more effective patient management.

In this regard, artificial intelligence (AI) is a promising tool (He *et al.* 2020). It is well established that deep learning approaches perform well on pattern recognition tasks in imaging data (Ting *et al.* 2019). Numerous approaches have been explored including classification of disease activity in AMD on optical coherence tomography (OCT) scans (Lee, Baughman & Lee 2017, Hwang *et al.* 2019, Motozawa *et al.* 2019, Potapenko *et al.* 2021), detection of referable disease on fundus images (Burlina *et al.* 2018, Bhuiyan *et al.* 2021) and quantifying pathological fluid by segmentation of OCT images (De Fauw *et al.* 2018, Schlegl *et al.* 2018, Gao *et al.* 2019).

Attempts have been made to enhance performance, for example by including all three dimensions of an OCT volume (Li *et al.* 2019) or using advanced model architectures like a capsule network (Tsuji *et al.* 2020). However, an area currently left largely unexplored is the temporal change in imaging. As clinicians often assess changes over time during patient evaluation, at minimum comparing the previous and current findings, the temporal dimension might offer additional information needed to detect disease activity.

Despite encouraging performance *in silico*, many of the published approaches are difficult to directly implement in a clinical environment. Most are restricted to a single input data type, which differs significantly from the much broader scope of data a clinician routinely evaluates during patient follow-up. Visual acuity, OCT scans and fundus photographs for both the current and previous visits will often be used to arrive at treatment decisions (Brown & Regillo 2007). Further, concepts like chronic oedema, unexpected changes in visual acuity and discharge to primary care ophthalmologists are not accounted for by most models, limiting their usefulness in any real-world scenario.

In this study, we present a novel design for a follow-up system of AMD

patients based on a combination of temporally aware AI and deterministic logic. This framework is tailored to autonomously suggest patient treatment in accordance with observe-and-plan (O&P) regimen, and handles several advanced concepts such as regimen compliance, chronicity and decision uncertainty. Through validation that closely resembles a clinical environment, we show the system to be safe and highly compliant with multi-expert consensus on treatment strategy in a prospectively collected data set. Such clinically oriented validation has not been previously attempted in the published literature and might lower the barriers to a future real-world implementation of the system.

Materials

Two data sets were used: A retrospective cohort for training and hyperparameter tuning of the AI model designed to detect disease activity and a prospective case cohort to evaluate the comprehensive follow-up system that comprised the AI model augmented with additional deterministic logic.

Retrospective training data set

The retrospective cohort was based on the previously described OCT image

data set in Potapenko *et al.* (2021) collected at the Department of Ophthalmology, Rigshospitalet, Copenhagen. The data were split into training, tuning and three validation sets (termed internal, external A and external B) that differed in the labelling procedure (Table 1).

As described in the original article, training, tuning and internal validation sets used labels derived from clinical decisions made according to pro-re-nata (PRN) regimen. If treatment had been prescribed, it was assumed that oedema (defined as pathological fluid either intra- or subretinally) was present on the OCT scan, otherwise it was assumed that oedema was not present. External validation sets A and B were manually re-graded by one or three graders, respectively, to denote whether the scan contained oedema or not.

Several modifications were made to the data sets in the current study. Fundus images associated with the OCT scans were kept. Temporal information was preserved by pairing each OCT scan with the scan taken during the previous visit. Patients in the prospective cohort (described in the next subsection) have been removed from the retrospectively collected data to ensure a complete separation of training and validation data sets. Radial OCT scans were included in

Table 1. Overview over data sets used.

Dataset	Period	Type	Regimen	Patients (n)	Scans (n)	Eyes (n)	Used labels
Training data sets	Jun 2007–Jun 2018	Retrospective	PRN				
Training set				4.898	84.489	6.194	Presence of oedema (derived from treatment decision)
Tuning set				612	10.107	750	Presence of oedema (derived from treatment decision)
Retrospective validation sets	Jun 2007–Jun 2018	Retrospective	PRN				
Internal validation set				612	10.411	779	Presence of oedema (derived from treatment decision)
External validation set A				215	1.446	265	Oedema presence (re-graded by a single expert)
External validation set B				187	187	187	Oedema presence (re-graded, three expert consensus)
Prospective validation set	Sep 2020–Feb 2021	Prospective	O&P				Two sets of labels available for all data
Eyes in active treatment (active)				100	100	100	Treatment decision (made in clinic)
Eyes observed w/o treatment (passive)				100	100	100	Treatment decision (re-graded, three expert consensus)

The above table gives an overview over the used data sets along with the following information: name of the data set; whether the data collection was prospective or retrospective; the period from which the data was collected; the treatment regimen that was used at the time; number of patients, eyes and OCT scans in each data set; how each scan was labelled. Please see the Materials section for description of the labelling procedures. Data used for regimen compliance calculation is omitted for clarity.

training and tuning sets in addition to volume scans from the original study, significantly increasing available data.

The modified data set included in total 84 489 scans (4898 patients) in the training set and 10 107 scans (612 patients) in the tuning set. The internal validation set contained 10 411 scans (612 patients); external validation sets A and B contained 1446 scans (215 patients) and 187 scans (187 patients) respectively.

Prospective expert-graded gold standard data set

Informed consent was prospectively collected from 230 patients during the period from 01.09.2020 to 01.02.2021 at the Department of Ophthalmology, Rigshospitalet, Copenhagen. From these, a selection of 200 eyes was made: 100 eyes in active treatment, and 100 eyes not receiving active treatment – either undergoing routine follow-up as a contralateral treatment-naïve eye or as a previously treated eye that is not showing signs of CNV activity. Eyes that were not treated in compliance with the department's regimen as defined below, were not eligible for inclusion. A single follow-up examination within the prospective period was selected. Historical data were recorded for every examination up to and including the selected visit, comprising visual acuity, treatment decision (number of prescribed injections with interval and anti-VEGF agent, or length of observation if no intravitreal injections were prescribed), treatments administered, OCT and fundus images and status of fellow eye (last treatment decision, if any, and visual acuity).

Treatment regimen and compliance

The Department of Ophthalmology follows a modified O&P regimen, originally described by Parvin *et al.* (2017). Patients' treatment is administered in pre-defined intensities, adjusted according to disease activity following rules set out by a drug-specific flow-chart (Fig. S1). However, a clinician might, on a case-by-case basis, choose to deviate from the pre-defined department guidelines.

For the purposes of the current study, compliance to treatment regime at any given consultation was operationally defined by two criteria. First,

the clinician-prescribed treatment during the last follow-up must follow the O&P treatment guidelines as described above. Second, the patient must have received the exact number of injections prescribed, at intervals that do not deviate more than 25% from the prescription.

For the purpose of evaluating compliance with the O&P regimen in the clinic, data on prescribed and administered treatments between 01.09.2018 and 01.09.2019 were used, comprising 11 215 follow-up consultations.

Ethics approval

This study was approved by the regional Data Protection Agency (jr. nr. P-2019-726) and the Danish Patient Safety Authority (filing nr. 3–3013-3214/1).

Methods

The goal of the current study was to construct and evaluate the safety of a comprehensive autonomous system for the treatment of AMD patients during follow-up, built around AI-based disease activity detection.

The initial steps were a continuation of our previously published efforts to train a deep learning model to recognize oedema on OCT scans (Potapenko *et al.* 2021). We trained an improved temporally aware model on the previously published retrospectively collected data set, where labels were derived from PRN compliant treatment decisions. Validation was performed on the same three validation cohorts, including two smaller sets that were manually re-graded by experts to denote the presence of oedema.

Our department has since transitioned to the O&P regimen. By using deterministic logic to augment the AI model, the output could be presented as treatment decisions compliant with the new regimen. We evaluated this system in a prospectively collected data set of patient cases from the retina clinic to ascertain its safety in a scenario closest possible to a real-world clinical setting. Original treatment decisions from the retina clinic and the AI-suggested treatment decisions were compared with the gold standard treatment decision in each case, defined by consensus between three retina experts.

An overview of the training and validation set-up is summarized in

Fig. 1, with a more complete overview of the data sets presented in Table 1. The process is described in more detail below.

AI-based system for follow-up of AMD patients

The AI system for follow-up of AMD patients was designed to make treatment decisions by evaluating clinical and imaging data from the current and the previous examinations. The system consists of two main components: (i) a deep learning-based model designed to detect CNV activity on OCT scans, and (ii) a layer of deterministic logic that defines how the output of this model should impact the management and treatment of the patient. Such functional division is analogous to the decision-making process in clinical practice: initially, the presence of CNV activity is established by assessing clinical and imaging findings; then, by applying clinical guidelines (*i.e.* those outlined by a regimen), the clinician determines further treatment.

AI-based temporal detection of disease activity

A deep learning model was constructed to detect CNV activity on OCT scans. Two model architectures with different inputs were tested: (i) a model that analysed current and previous OCT scans, and (ii) a model that analysed current and previous OCT scans along with the current fundus photograph. A brief overview of the technical implementation can be found below.

Previously reported model architecture (Potapenko *et al.* 2021) was used as a basis, with the following alterations: (a) all B-scans were used instead of the central six, connected by a max-sum-pooling layer, (b) a spatial dropout layer with rate of 0.2 was introduced after each convolution layer, and (c) input resolution of slices was increased to 384×384 (Fig. S2A and B). The model was trained using the training and tuning data sets as defined above. Training was otherwise conducted as described in Potapenko *et al.* (2021).

To implement a temporal deep learning network, transfer learning was used. After training the above model, the final fully connected layer was discarded, and the weights of all remaining layers were frozen. Each of

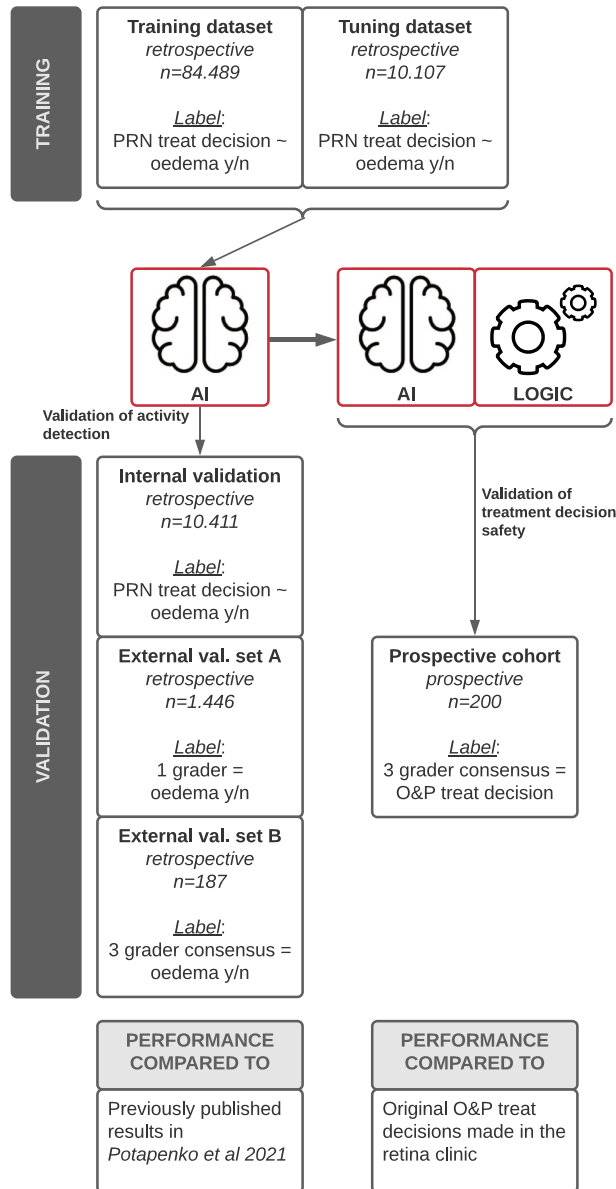


Fig. 1. Stages of development and data utilization. This figure provides an overview of how the various sources of data were used during the training and validation processes. Data sets used for training, tuning and validation are shown, each with the following information displayed: number of cases, whether the data were retrospectively or prospectively collected, and labelling information. For the latter, the source of the labelling and what the labels denote is given. The source of the labelling is accompanied by either an equality sign ('=') when expert re-grading is used, or a tilde ('~') if labels are derived indirectly from PRN treatment decisions (refer to the Materials section for more information). The figure differentiates between the validation of the AI component alone (red pictogram 'AI') and the validation of the entire system (i.e. the AI component together with the deterministic logic represented by the red pictogram 'Logic'). Finally, the figure states which sources the current performance metrics were compared with. For a more detailed description of the data sets, see Table 1.

the two consecutive OCT scans was processed by the pre-trained layers, with the output of both combined using a new trainable fully connected layer (Fig. S2C). This temporal model could then be trained using the training and tuning data sets as before but now with two consecutive OCT scans as

inputs. The training process was otherwise identical to the single scan model above.

To evaluate whether including the fundus image from the current examination as an input could increase the performance further, a second version of the temporal deep learning network

was created. This was done by adding a third model alongside the two OCT scan models, feeding forward to the final fully connected layer (Fig. S2D). The added network was structurally identical to the one described for a single OCT scan and was fully trainable. To compare performance, percentile bootstrap method was used to get 95% confidence intervals for all performance metrics.

Decision thresholds for the temporal model both with and without fundus image input were defined by inspecting their reliability curves (Fig. S3). Confidence thresholds were chosen for negative (disease activity not present) and positive (disease activity present) classifications. If model output was numerically between these, the prediction was considered unreliable.

Deterministic logic

The deterministic logic layer integrates the AI model's prediction of disease activity with clinical parameters to either make an autonomous treatment decision (output 9 below) or ask for a second opinion from a human ophthalmologist (output 1 through 8). When asking for a second opinion, the system will attempt to suggest an action (output 2, 3 and 5 through 8) or give a reason for being unable to make an autonomous decision (output 1 and 4). All outputs are designed to be easily understandable by an ophthalmologist with no prior experience with the system or knowledge of its internal workings.

All possible outputs from the system are listed below; the corresponding logic diagram can be seen in an abridged version in Fig. 2, and in full in Fig. S4.

- 1 Inclusion criteria not met. Data are missing, the treatment prescribed is non-compliant with the regimen or the patient has not been compliant with the prescribed treatment, as per rules defined above.
- 2 Intraocular pressure high. If measured above 25 mmHg, an ophthalmologist needs to be consulted for applanation tonometry and further evaluation.
- 3 Visual acuity is below treatment threshold. If the best-corrected visual activity (BCVA) is below 0.1 decimal, an ophthalmologist is needed to assess whether treatment should be terminated.

- 4 AI model is uncertain. If AI model output does not fall within the confidence thresholds defined above, a second opinion from a human ophthalmologist is required.
- 5 New activity in an inactive CNV. Provided that no disease activity was present during the last follow-up, and no injections were prescribed or given, newly detected re-activation of a CNV requires a human ophthalmologist to set the initial treatment frequency.
- 6 Treatment-resistant oedema. Disease activity is present despite the last two prescriptions being the most intensive on the treatment chart without improvement in BCVA (less than 0.2 decimal). Re-evaluation of treatment strategy or revision of the diagnosis might be required.
- 7 Potential concurrent eye disease. If BCVA has decreased at least 0.2 decimal since the last visit, but no CNV activity is detected, examination by an ophthalmologist is warranted to rule out concurrent ocular comorbidity.
- 8 Discharge to primary care ophthalmologist. The patient can be followed in the primary sector if neither eye has shown CNV activity or received intravitreal injections during the last 6 months.
- 9 Treatment suggestion. If none of the above exceptions occur, the current position in the O&P treatment diagram is found and – based on the AI-determined presence of CNV activity – the appropriate treatment option from the diagram is chosen.

Graders

Three graders (FU, TI and JN) with 20, 12 and 8 years of experience in treating AMD patients, respectively, independently evaluated the examinations in the prospective cohort using a specially designed grading tool (Fig. S5). For each case, the graders could examine a timeline of all previous examinations including visual acuity, previous treatment decisions (number of prescribed injections with interval and anti-VEGF agent, or length of observation if no intravitreal injections were prescribed), treatments given, OCT and fundus images, and status of the fellow eye (last treatment decision, if any, and visual acuity). After reviewing this information, each grader was asked to make a treatment decision.

Consensus was defined as either a majority decision where at least two of the graders had submitted identical decisions, or in cases where all three decisions were different, a re-evaluation of the case in a plenary meeting to establish a decision accepted by at least two of the three graders.

AI system performance evaluation

Three *a priori* defined primary endpoints for this study were (1) the system's safety, *that is* agreement of AI treatment decisions with expert consensus compared with agreement between decisions made in the clinic with expert consensus, (2) the system's operational potential, defined as the proportion of regimen-compliant follow-ups that can be evaluated without human intervention and (3) potential for overall automation, defined as the percentage of follow-ups where an autonomous decision can be taken, *that is* regimen compliance in the department multiplied by operational potential.

Regimen compliance

To evaluate to which degree regimen guidelines are followed in clinical practice, compliance criteria previously described in the Materials section were applied to each treatment decision in the retrospective data. If the prescribed treatment both followed department guidelines and was administered as prescribed, the decision was classified as compliant. If a clinician chose not to follow O&P guidelines when making treatment decision, non-compliance was classified as prescription-related (*i.e.* prescription did not adhere to the regimen). If the prescription was made in accordance with department guidelines but was not administered as prescribed (*i.e.* timing or number of injections given did not match the prescribed), it was classified as patient-related.

Decision safety classification

Decision safety classification was based on whether the prescribed treatment was identical with the expert consensus in terms of frequency, number and type of injections, or time to follow-up in case no intravitreal injections were required. If identical, the treatment was considered safe; otherwise, it was considered unsafe. Two special cases

were considered separately. If the gold standard was to discharge the patient to the primary care ophthalmologist, but a decision was made to keep follow-ups at the hospital, the decision was classified as safe, as it does not pose a risk to the patient. If the gold standard was to observe chronic oedema without treatment, but a decision was made to continue intravitreal injections, the decision was classified as unsafe, as it could potentially expose the patient to unnecessary risk.

The AI system's request for a second opinion is safe by definition, as the final decision is taken by a human ophthalmologist. The system will nonetheless suggest further treatment in some of these cases; these can be similarly classified as potentially safe or potentially unsafe based on the rules described above.

Results

Regimen compliance

Regimen compliance in the retina clinic was found to be 81.6%. Prescription-related non-compliance was found in 9.4% of eyes, with the remaining 9.0% being due to patient-related non-compliance.

Expert grading

Unanimous agreement was found in 46% of cases, higher in the passively observed eyes (56%) than in actively treated eyes (36%, $p = 0.007$). No two graders were markedly more in agreement than the other (number of cases with agreement between JF and TI, JF and JN, TI and JN, were 56%, 58% and 63.5% respectively). In 11.5% ($n = 23$) of the cases agreement was not present after individual grading, and consensus was instead reached during a plenary meeting. Of these, disagreements were on following topics: number and interval of injections ($n = 8$), presence of chronic oedema ($n = 6$), time to follow-up appointment ($n = 6$), whether the patient should be discharged ($n = 2$) and whether a differential diagnosis should be considered ($n = 1$). Graders chose to deviate from the clinic's treatment regimen guidelines in 2.5% of cases ($n = 5$), mostly on suspicion of chronic or treatment-refractory oedema ($n = 3$).

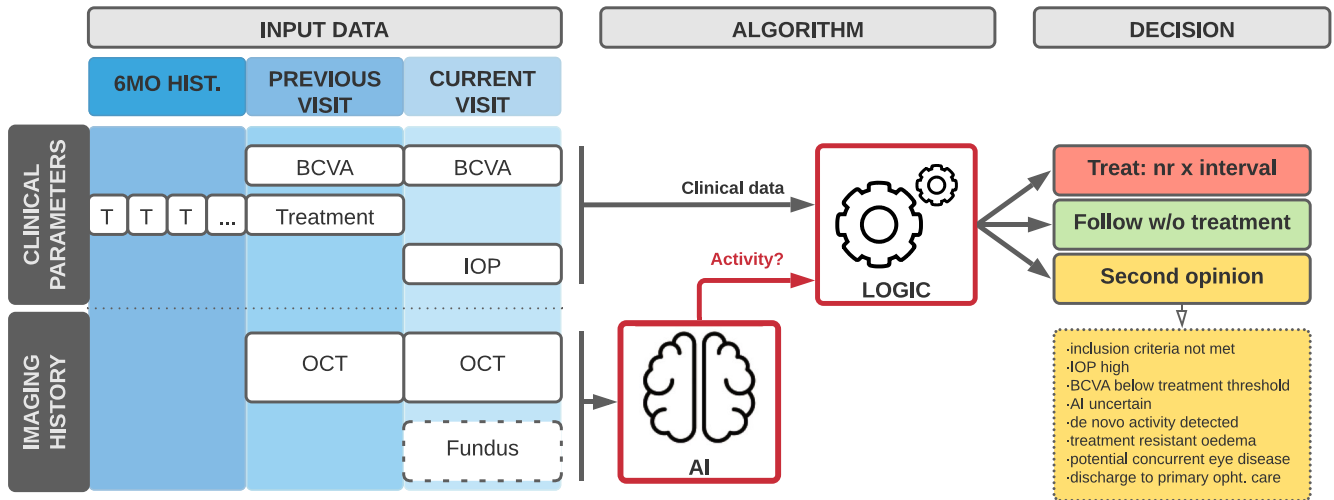


Fig. 2. AI system component overview. This figure provides an overview of the input and output of the AI-based system, along with a simplified schematic of the internal components (the AI model and the deterministic logic). On the left, inputs are shown: clinical data from the current and previous visit, along with treatment history of the prior 6 months, are directly processed by the deterministic logic component; OCT scans from the previous and current visit are input first into the AI model (along with the fundus photograph for the model that supports it), and then the activity score is passed into the deterministic logic component. After processing, the deterministic logic will output either an autonomous treatment decision (number of injections with a given interval or follow without treatment) or a request for second opinion (elaborated by a reason or a suggestion for further actions; list of possibilities shown on the right in yellow). For a more comprehensive overview of the deterministic logic component, see Fig. S1. T – treatment, IOP – intraocular pressure, BCVA – best-corrected visual acuity, OCT – optic coherence tomography scan.

Performance evaluation of the temporal AI model

Both new models with and without fundus photographs showed several percentage points higher AUC than the previously published model (0.900 (95% CI 0.894–0.906) for both new models versus 0.857 (95% CI 0.846–0.867) for the old model in the internal validation data set), with the same order of improvement in sensitivity (0.836 (95% CI 0.827–0.845) and 0.837 (95% CI 0.828–0.846) versus 0.762 (95% CI 0.746–0.778) respectively; detailed performance statistics can be found in Table S1). Trends towards improved performance could be seen for some metrics (sensitivity and accuracy) in the external validation sets, but 95% confidence intervals overlapped. The new temporal model that incorporated fundus images failed to improve performance over the model without the fundus photographs but increased architectural complexity. It was therefore decided to exclude it from further analyses.

Performance evaluation of the comprehensive AI-based system for follow-up of AMD patients

When validating against expert consensus, the comprehensive AI system made safe autonomous decisions in 67%

($n = 134$) and unsafe decisions in 6% ($n = 12$) of eyes, while asking for second opinion in 27% ($n = 54$) of cases (Table 2). Upon inspection, unsafe decisions were due to presumption of CNV activity where experts did not find any (*i.e.* false positives; $n = 8$), regular treatment where experts decided on different treatment due to chronicity ($n = 2$), lack of detected activity where experts found de-novo CNV activity (*i.e.* false negatives; $n = 1$), and regular treatment where experts wanted further imaging to rule out differential diagnosis ($n = 1$). A second opinion was mainly requested due to the AI model not reaching confidence level required for an autonomous decision (46.3%, $n = 25$). For the remainder of the second opinion requests ($n = 29$), treatment suggestions supplied by the AI-based system were correct in 79.3% ($n = 23$) of the cases.

The AI decisions were compared with the decisions made in the retina clinic in terms of their conformity with the expert consensus. Cases where the AI model gave prediction above confidence threshold (autonomous or non-autonomous) were considered ($n = 176$). The AI system made the same decisions as the expert consensus (*i.e.* safe decisions) in 89.2% ($n = 157$) of the cases, while the retina clinic

agreement rate with expert consensus was 85.8% ($n = 151$; $p = 0.42$). If only cases where the AI system took an autonomous decision were considered ($n = 146$), 91.8% ($n = 134$) were in agreement with the expert consensus, while in the retina clinic, 87.7% ($n = 128$) of the decisions agreed with the consensus ($p = 0.33$).

Interestingly, when only considering the cases that needed to be discussed at the consensus meeting ($n = 23$), almost all decisions the AI-based system made were second opinions ($n = 21$; 20 of which were safe suggestions), one autonomous decision was safe, and one was unsafe. In comparison, the retina clinic decisions were split approximately in half by safe and unsafe decisions ($n = 12$ and $n = 11$ respectively).

If the AI-based system could theoretically be tested on all AMD patients followed at the department, the above results need to be corrected for the fact that 18.4% of the patients are non-compliant with the regimen and are thus by design excluded from consideration by the algorithm. In absolute terms, the results would then translate into 22.0% of patients being sent to a second opinion by a human ophthalmologist, 54.7% being given a correct treatment and 4.9% being treated incorrectly.

Table 2. Safety classification of decisions in the prospective validation data set made by the AI system and the clinical staff at the retina clinic.

	n		%	
(A)				
Autonomous decisions	146		73.0	
Safe decision	134		67.0	
Unsafe decision	12		6.0	
Second opinion	54		27.0	
Safe suggestion	23		11.5	
Unsafe suggestion	6		3.0	
AI pred. below confidence threshold	25		12.5	
Total	200			
(B)				
Safe decision	151		75.50	
Unsafe decision	49		24.50	
Total	200			
	AI		Retina clinic	
	n	%	n	%
(C)				
Safe decision	134	91.8	128	87.7
Unsafe decision	12	8.2	18	12.3
Total	146		146	

Briefly, if the decision was identical with expert consensus decision, it was considered safe, otherwise it was considered unsafe. See the Methods section for more detailed description. (A) Safety classification of treatment decisions made by the AI system for the entire prospective data set ($n = 200$); (B) Safety classification of treatment decisions made in the retina clinic for the entire prospective data set ($n = 200$); (C) Safety profile comparison between treatment decisions made in the retina clinic and by the AI system among the cases where the AI system made an autonomous decision only ($n = 146$).

Discussion

The continued rise in the number of AMD patients expected during the coming decade (Potapenko & la Cour 2021) will pose a challenge to public ophthalmology services. Increased staffing is unlikely to be a viable solution on its own, prompting the need for novel approaches. In this study, we propose the first ever comprehensive system for autonomous follow-up of AMD patients that might fulfil this role.

The system’s most basic component is its temporal AI model that detects disease activity on OCT images. To train such AI classifier requires data labelled with the presence of oedema, but with a data set of over 100 000 OCT scans, manual labelling is impractical.

Labels that are easiest accessible can be derived from treatment decisions. These, however, are heavily influenced by the treatment regimen used during the time the data were collected, and are not necessarily applicable in a setting of a different regimen. We have previously shown that a well-performing AI model for the detection of oedema can be trained on proxy labels derived from PRN treatment decisions (Potapenko *et al.* 2021). In the current study, this model was further improved by including information from the previous examination, analogously to what a clinician might do during patient evaluation. A layer of deterministic logic was added so that the output of the AI model (*i.e.* whether CNV activity is present) could be translated into a treatment decision compatible with the O&P regimen currently in use at our department. Information about visual acuity, important in determining if further treatment is futile, was also incorporated in the logic layer.

To gauge the performance of the combined system of AI and deterministic logic, a clinically oriented validation was performed. We prospectively collected a data set of 200 real-world AMD follow-ups, including the treatment decision made by clinicians at our department. Each case was then re-evaluated by three experienced retina specialists who had access to full clinical details and imaging history. Agreement between the AI system’s decisions and the expert consensus was compared with the original treatment decisions taken in the retina clinic, providing information about the system’s safety. Finally, we attempted to determine the potential proportion of AMD patients that could be autonomously followed by the proposed AI system.

Temporally aware AI system architecture

The changes in the previously published AI model architecture (Potapenko *et al.* 2021), resulted in a substantial performance improvement. The noticeable false-negative rate of the previous model, hypothesized to be related to undetected small quantities of oedema, has been reduced and is only 0.5% in the prospective data set. This presumably relates to the new model’s higher input resolution and the addition of the temporal dimension.

The former is supported by similar results for classification tasks in radiology, where a larger input size can improve performance (Sabottke & Spieler 2020). The temporal approach is novel and has never been attempted for detection of neovascular AMD activity before. Some research has, however, suggested that recurrent neural networks – that by design operate on temporal data – might be of value, especially in progression analysis (Jiang *et al.* 2018) but also disease detection (Gheisari *et al.* 2021).

Performance evaluation showed significant improvement in the data set with labels derived from clinical decisions, a process where temporal data is actively used by the clinician. Improvement was much less evident in the data sets relabelled according to the presence of oedema based on isolated scans only. This could relate to the current model incorporating additional information not contained within a single scan, although the far smaller size of the external validation sets undoubtedly plays a role.

The addition of fundus images had little effect on the model’s performance. Part of the explanation might be that almost 90% of haemorrhages visible on a dilated fundus examination result in structural changes on OCT, and those that do not, might not require treatment (Patel *et al.* 2020). Moreover, the current false negative rate is already low, presumably further limiting the usefulness of fundus photography as a training input; a fundus examination is more useful in detecting CNV activity than ruling it out.

Labelling incongruence

Label incongruency has been covered in some length in our previous work (Potapenko *et al.* 2021). In brief, a PRN treatment decision to treat or not to treat usually encapsulates evaluation of more than just oedema. Other considerations may also play a role, including haemorrhage on fundus photography, chronicity, patient preferences, compliance to prescribed treatment and a clinician’s decision not to follow guidelines. Thus, it is not immediately clear that these labels are suited as surrogates for an oedema-based classification of OCT images. Nonetheless, we demonstrate an AI model for oedema detection that

performs well despite being trained on data with labels based on PRN treatment decisions.

Temporality introduces additional input heterogeneity (Fig. S6). Treatment status between the previous and the current OCT is not included in the classifier. Time between examinations, number of injections, injection intervals and anti-VEGF agent used can thus not be taken into account by the AI model. Further complexity is introduced by a number of other factors, for example individual sensitivity to treatment and quantitative changes in oedema over time. It is therefore not obvious that including the temporal dimension would be advantageous. As the performance nonetheless increased, some structural elements independent of the aforementioned factors must be present on the previous OCT scans that aid in oedema detection.

Additional label complexity is present in the O&P data, since treatment during de-intensification can be given even if no oedema is observed (Fig. S2), while this cannot occur during PRN-compliant treatment (Fig. S6). The difference, however, appears not to have influenced oedema detection accuracy in the prospective validation cohort. Conversely, this shows that PRN-derived labels are more suitable for training oedema detection than O&P labels, as they, although not equivalent to oedema, minimize input complexity.

Gold standard gradings

There are few studies that compare independent expert grading in a similar setting. We have previously reported agreement between three experts on the presence of oedema on OCT scans to be 76.4% (Potapenko *et al.* 2021). This is comparable to previously reported findings, although, depending on which features were assessed, inter-rater agreement can vary widely (Chandra *et al.* 2021, Müller *et al.* 2021). A clinical management decision for AMD patients, despite being largely based on the presence of oedema, is more nuanced, and a lower consensus would be expected. Even having taken this added complexity into consideration, unanimous agreement in less than half of the cases must be considered low. A significantly higher agreement was seen in the eyes not being actively treated, which is to be expected as complexity is

presumably less in these mostly dry retinas. The consensus meeting showed that the most common diverging opinions were on what should be considered activity and chronic fluid. This is unsurprising, as the topic of refractory oedema is known to be complex, with some advocating not to treat residual subretinal fluid in certain cases (Bhavsar & Freund 2014, Jang *et al.* 2015).

Operational potential and regimen compliance

Differing clinical opinions also manifest in non-compliance with regimen guidelines. We found significant non-adherence to prescription guidelines in the retina clinic (9%), which was almost fourfold higher than the experts' (2.5%). Discrepancies are perhaps even more pronounced at different hospital departments: Age-adjusted rates of anti-VEGF injections among AMD patients varies greatly between the Danish Regions, from 3% in Northern Jutland to 6% in the Capital and Southern Regions (Vittrup 2019). These substantial differences, though doubtlessly affected by a number of factors, raise the possibility of patients receiving disparate levels of care in different regions. An advantage of a common autonomous follow-up system is that it always provides regimen-compliant treatment suggestions. This might help deliver equal quality of care to all patients independent of clinical environment and geographical location.

As a best-case scenario, the automation potential reported in this study would imply an almost two-thirds reduction in the number of eyes seen by ophthalmologists for AMD follow-up at the retina clinic. A further improvement in operational potential can occur if implementation of the system reduces non-compliance, consequently increasing the number of patients that can be controlled autonomously over time. The system's flexibility allows patients to be freely moved in and out of automatic follow-up, even after a period of non-compliance, provided the last consultation and treatment have both been compliant in terms of prescription and execution. This means that all patients are eligible to be followed up autonomously at least for a proportion of the follow-ups they attend, provided they are not consistently treated non-compliantly.

Safety and future potential of AI-based system for AMD patient follow-up

We report for the first time on safety of a novel AI-based system for management of AMD patients, validated in a prospective cohort of clinical cases. Few real-world clinical implementations of AI systems have been adopted to date, and only two systems have received approval by the United States Food and Drug Administration (Abràmoff *et al.* 2018, Ipp *et al.* 2021). Both the rationale behind the clinically oriented system design and validation procedure reported in the current study are intended to shorten the path to clinical implementation.

Conformity of autonomous decisions to expert consensus was on par with a large retina clinic. Two key design aspects contributed to the system's safety: relegating the responsibility of making the primary diagnosis to human ophthalmologists, and allowing for second opinion requests. The former significantly reduces (although does not entirely eliminate) the risk of missing comorbidities and rare ocular conditions, which would otherwise pose a challenge for an AI system due to limited training data (Ting *et al.* 2019). Consequently, the system can be considered as a modality for treatment titration *after* the indication has been established – not as an autonomous diagnosis and treatment modality, potentially easing the legal approval process.

The second opinion requests were implemented to ensure the safety of autonomous decisions by avoiding complex cases and major treatment changes. The validation appears to support that this worked as intended; in eyes where a consensus meeting was needed to achieve agreement between the experts (*i.e.* presumably the most complex cases), the AI system almost exclusively asked for second opinions. Both among these cases, and second opinion requests in general, almost half received a suggestion for treatment that matched the expert consensus.

The proposed modular design with a separate disease activity detecting AI model and a guideline directed non-AI component makes the system adaptable to a variety of settings and departments. Treatment regimens, such as PRN or other variants of treat-and-extend, can be implemented by altering

the deterministic logic without the time and data required for re-training of the AI model. System parameters, such as the inclusion criteria and confidence thresholds, can be easily adjusted *ad hoc* after system deployment. This gives granular control over how many patients are managed by the system and the number of second opinion requests to human staff.

A large amount of research has been dedicated to improving AI's explainability. Our group, among others, has shown that methods such as class activation map (CAM) can be used to extract meaningful clinical information about parts of OCT scans critical to AI decisions, although this approach is not without significant limitations (Potapenko *et al.* 2021). Contrary to this, deterministic logic used in our system is more transparent and easily understandable: It consists of a series of explicitly defined conditions and outcomes that can be understood and followed intuitively (Fig. 2) without the need for a complex interpretation mechanism.

A future implementation of this system in coordination with ophthalmology care providers in the primary sector seems to have several advantages. At the hospital level, this could mean a significant reduction in the number of follow-ups performed by ophthalmologists. Human resources can thus be re-allocated to other tasks, for example reducing delays in the initial AMD diagnosis and treatment, which is known to be of major significance for patient outcomes (Ho *et al.* 2017). Alternatively, patients can be followed locally using the system deployed by a primary care provider, only visiting the hospital for injections. This would allow for shorter patient travels, which is both beneficial for the many elderly with limited mobility and significant comorbidities and reduces the environmental footprint. Finally, a low rate of false negatives indicates that the system is well suited for long-term screening of patients with quiescent CNVs outside of the hospital environment.

Limitations

This study demonstrates usability of the AI-based system on real-world patient cases, closely resembling follow-up consultations in a retina

clinic. However, the current study is still done *in silico*, with the logical next step being an evaluation of a live implementation alongside the clinicians and validation at other institutions. This might provide further performance metrics and more granular information on whether any critical information is lost by omitting physical examination and direct communication with the patient. Importantly, it may elucidate the system's acceptance among clinicians and patients. This aspect has been explored in the literature before (Prah & Enright 2017), but few studies have tested it in clinical practice. Lacking support from personnel or rejection by the patients might be a major barrier to the system's implementation.

The structure of the current study did not allow estimation of how many patients cannot be followed automatically due to fellow eye needing either a second opinion or not fulfilling the inclusion criteria. This may possibly have an impact on real-world automation potential.

The architecture of the AI model could potentially be further enhanced. Quantification of oedema may be helpful in a proportion of the decisions made, for example in terms of response to treatment and chronicity (Schmidt-Erfurth *et al.* 2020, Schmidt-Erfurth *et al.* 2021) – although this would arguably significantly increase the complexity of both the AI model and the deterministic logic. Inclusion of treatment administered between the current and the previous OCT scans would likely improve performance, but not enough data could be gathered from the period when the department used the O&P regimen. Training on data from the PRN regimen period is not possible due to differing implication of treatment status as previously described: treatment is administered only due to signs of CNV activity in PRN regimen, while dry retinas can also be treated according to the O&P regimen.

Finally, it is possible, that some regimens require other parameters that might not be as easy to encode deterministically using the current algorithm structure (for example concepts like 'irreversible foveal damage'), and additional changes would be required. In any case, a re-coded deterministic logic component will require additional validation.

Conclusions

This is the first study to demonstrate the feasibility and safety of a comprehensive temporally aware AI-based system for follow-up of AMD patients. By comparing its performance to a unique expert consensus data set containing 200 real-world cases, we were able to demonstrate performance on par with the decisions made in the retina clinic, with especially low rates of false-positive classification of CNV activity. Over half of the eyes with AMD could be followed autonomously without additional risk to the patient. The current results are encouraging; however, a live implementation is needed to establish real-life performance. Even if not deployed to its full potential, the proposed algorithm could greatly ease the pressure on the public ophthalmology departments from an increasing number of AMD patients and be a part of an efficient system for follow-up and treatment in concert with the primary sector services.

References

- Abramoff MD, Lavin PT, Birch M, Shah N & Folk JC (2018): Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med* **1**: 1–8.
- Baek SK, Kim JH, Kim JW & Kim CG (2019): Increase in the population of patients with neovascular age-related macular degeneration who underwent long-term active treatment. *Sci Rep* **9**: 12364.
- Bhavsar KV & Freund KB (2014): Retention of good visual acuity in eyes with neovascular age-related macular degeneration and chronic refractory subfoveal subretinal fluid. *Saudi J Ophthalmol* **28**: 129–133.
- Bhuiyan A, Govindaiah A, Alauddin S, Otero-Marquez O & Smith RT (2021): Combined automated screening for age-related macular degeneration and diabetic retinopathy in primary care settings. *Ann Eye Sci* **6**: 12.
- Brown DM & Regillo CD (2007): Anti-VEGF agents in the treatment of neovascular age-related macular degeneration: applying clinical trial results to the treatment of everyday patients. *Am J Ophthalmol* **144**: 627–637.
- Brynskov T, Munch IC, Larsen TM, Erngaard L & Sørensen TL (2020): Real-world 10-year experiences with intravitreal treatment with ranibizumab and aflibercept for neovascular age-related macular degeneration. *Acta Ophthalmol* **98**: 132–138.
- Burlina P, Joshi N, Pacheco KD, Freund DE, Kong J & Bressler NM (2018): Utility of deep learning methods for referability classification of age-related macular degeneration. *JAMA Ophthalmol* **136**: 1305–1307.

- Chandra S, Rasheed R, Sen P, Menon D & Sivaprasad S (2021): Inter-rater reliability for diagnosis of geographic atrophy using spectral domain OCT in age-related macular degeneration. *Eye* **2021**: 1–6.
- De Fauw J, Ledsam JR, Romera-Paredes B et al. (2018): Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med* **24**: 1342–1350.
- Gao K, Niu S, Ji Z et al. (2019): Double-branched and area-constraint fully convolutional networks for automated serous retinal detachment segmentation in SD-OCT images. *Comput Methods Programs Biomed* **176**: 69–80.
- Gheisari S, Shariflou S, Phu J, Kennedy PJ, Agar A, Kalloniatis M & Golzan SM (2021): A combined convolutional and recurrent neural network for enhanced glaucoma detection. *Sci Rep* **11**: 1–11.
- He M, Li Z, Liu C, Shi D & Tan Z (2020): Deployment of artificial intelligence in real-world practice: opportunity and challenge. *Asia-Pacific J Ophthalmol* **9**: 299–307.
- Ho AC, Albin TA, Brown DM, Boyer DS, Regillo CD & Heier JS (2017): The potential importance of detection of neovascular age-related macular degeneration when visual acuity is relatively good. *JAMA Ophthalmol* **135**: 268–273.
- Hwang DK, Hsu CC, Chang KJ et al. (2019): Artificial intelligence-based decision-making for age-related macular degeneration. *Theranostics* **9**: 232–245.
- Ipp E, Liljenquist D, Bode B et al. (2021): Pivotal evaluation of an artificial intelligence system for autonomous detection of referable and vision-threatening diabetic retinopathy. *JAMA Netw Open* **4**: e2134254.
- Jang L, Gianniou C, Ambresin A & Mantel I (2015): Refractory subretinal fluid in patients with neovascular age-related macular degeneration treated with intravitreal ranibizumab: visual acuity outcome. *Graefes Arch Clin Exp Ophthalmol* **253**: 1211–1216.
- Jiang J, Liu X, Liu L et al. (2018): Predicting the progression of ophthalmic disease based on slit-lamp images using a deep temporal sequence network. *PLoS One* **13**: e0201142.
- Lee CS, Baughman DM & Lee AY (2017): Deep learning is effective for classifying normal versus age-related macular degeneration OCT images. *Ophthalmol Retin* **1**: 322–327.
- Li M-X, Yu S-Q, Zhang W, Zhou H, Xu X, Qian T-W & Wan Y-J (2019): Segmentation of retinal fluid based on deep learning: application of three-dimensional fully convolutional neural networks in optical coherence tomography images. *Int J Ophthalmol* **12**: 1012–1020.
- Motozawa N, An G, Takagi S et al. (2019): Optical coherence tomography-based deep-learning models for classifying normal and age-related macular degeneration and exudative and non-exudative age-related macular degeneration changes. *Ophthalmol Ther* **8**: 527–539.
- Müller PL, Liefers B, Treis T et al. (2021): Reliability of retinal pathology quantification in age-related macular degeneration: implications for clinical trials and machine learning applications. *Transl Vis Sci Technol* **10**: 4.
- Papadopoulos Z (2020): Recent developments in the treatment of wet age-related macular degeneration. *Curr Med Sci* **40**: 851–857.
- Parvin P, Zola M, Dirani A, Ambresin A & Mantel I (2017): Two-year outcome of an observe-and-plan regimen for neovascular age-related macular degeneration treated with Aflibercept. *Graefes Arch Clin Exp Ophthalmol* **255**: 2127–2134.
- Patel Y, Miller DM, Fung AE, Hill LF & Rosenfeld PJ (2020): Are dilated fundus examinations needed for OCT-guided retreatment of exudative age-related macular degeneration? *Ophthalmol Retin* **4**: 141–147.
- Potapenko I, Kristensen M, Thiesson B et al. (2021): Detection of oedema on optical coherence tomography images using deep learning model trained on noisy clinical data. *Acta Ophthalmol* **100**: 103–110.
- Potapenko I & la Cour M (2021): Modelling and prognostication of growth in the number of patients treated for neovascular age-related macular degeneration. *Acta Ophthalmol* **99**: e1348–e1353.
- Prahl A & Enright RD (2017): Forgiving computers: the rise of automation and implications for counseling. *Couns Values* **62**: 144–158.
- Sabottke CF & Spieler BM (2020): The effect of image resolution on deep learning in radiography. *Radiol Artif Intel* **2**: e190015.
- Schlegl T, Waldstein SM, Bogunovic H et al. (2018): Fully automated detection and quantification of macular fluid in OCT using deep learning. *Ophthalmology* **125**: 549–558.
- Schmidt-Erfurth U, Reiter GS, Riedl S et al. (2021): AI-based monitoring of retinal fluid in disease activity and under therapy. *Prog Retin Eye Res* **86**: 100972.
- Schmidt-Erfurth U, Vogl WD, Jampol LM & Bogunović H (2020): Application of automated quantification of fluid volumes to anti-VEGF therapy of neovascular age-related macular degeneration. *Ophthalmology* **127**: 1211–1219.
- Ting DSW, Pasquale LR, Peng L et al. (2019): Artificial intelligence and deep learning in ophthalmology. *Br J Ophthalmol* **103**: 167–175.
- Tsuji T, Hirose Y, Fujimori K et al. (2020): Classification of optical coherence tomography images using a capsule network. *BMC Ophthalmol* **20**: 114.
- Vittrup M (2019): The Danish Patient Association for Prevention of Blindness, Fight for Sight; Personal communication (13.09.2019).

Received on November 28th, 2021.
Accepted on March 12th, 2022.

Correspondence:
Ivan Potapenko, MD, PhD
Department of Ophthalmology, Rigshospitalet
Valdemar Hansens Vej 1-23
2600 Glostrup
Copenhagen
Denmark
Telephone: +45 38 63 47 00
Fax: +45 38 63 46 69
Email: ivan.olegovich.potapenko@regionh.dk

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1 Schematic representation of the observe-and-plan regimen-based guidelines for (a) aflibercept and (b) ranibizumab.

Figure S2 Deep learning model architecture, represented at different stages of training.

Figure S3 Reliability curves.

Figure S4 Schematic representation of the deterministic logic that defines behaviour of the AI system.

Figure S5 Graphical user interface of the validation tool used by the experts to evaluate the prospective cohort of 200 cases.

Figure S6 An overview of possible combinations of input data and output labels.

Table S1 Performance metrics of the original model as reported in Potapenko et al. 2021, compared to the two new temporal model variants with and without inclusion of fundus image input.