Clinically relevant deep learning for detection and quantification of geographic atrophy from optical coherence tomography: a model development and external validation study

Gongyu Zhang, Dun Jack Fu*, Bart Liefers, Livia Faes, Sophie Glinton, Siegfried Wagner, Robbert Struyven, Nikolas Pontikos, Pearse A Keane, Konstantinos Balaskas*

Summary

Background Geographic atrophy is a major vision-threatening manifestation of age-related macular degeneration, one of the leading causes of blindness globally. Geographic atrophy has no proven treatment or method for easy detection. Rapid, reliable, and objective detection and quantification of geographic atrophy from optical coherence tomography (OCT) retinal scans is necessary for disease monitoring, prognostic research, and to serve as clinical endpoints for therapy development. To this end, we aimed to develop and validate a fully automated method to detect and quantify geographic atrophy from OCT.

Methods We did a deep-learning model development and external validation study on OCT retinal scans at Moorfields Eye Hospital Reading Centre and Clinical AI Hub (London, UK). A modified U-Net architecture was used to develop four distinct deep-learning models for segmentation of geographic atrophy and its constituent retinal features from OCT scans acquired with Heidelberg Spectralis. A manually segmented clinical dataset for model development comprised 5049 B-scans from 984 OCT volumes selected randomly from 399 eyes of 200 patients with geographic atrophy secondary to age-related macular degeneration, enrolled in a prospective, multicentre, phase 2 clinical trial for the treatment of geographic atrophy (FILLY study). Performance was externally validated on an independently recruited dataset from patients receiving routine care at Moorfields Eye Hospital (London, UK). The primary outcome was segmentation and classification agreement between deep-learning model geographic atrophy prediction and consensus of two independent expert graders on the external validation dataset.

Findings The external validation cohort included 884 B-scans from 192 OCT volumes taken from 192 eyes of 110 patients as part of real-life clinical care at Moorfields Eye Hospital between Jan 1, 2016, and Dec, 31, 2019 (mean age 78·3 years [SD 11·1], 58 [53%] women). The resultant geographic atrophy deep-learning model produced predictions similar to consensus human specialist grading on the external validation dataset (median Dice similarity coefficient [DSC] 0·96 [IQR 0·10]; intraclass correlation coefficient [ICC] 0·93) and outperformed agreement between human graders (DSC 0·80 [0·28]; ICC 0·79). Similarly, the three independent feature-specific deep-learning models could accurately segment each of the three constituent features of geographic atrophy: retinal pigment epithelium loss (median DSC 0·95 [IQR 0·15]), overlying photoreceptor degeneration (0·96 [0·12]), and hypertransmission (0·97 [0·07]) in the external validation dataset versus consensus grading.

Interpretation We present a fully developed and validated deep-learning composite model for segmentation of geographic atrophy and its subtypes that achieves performance at a similar level to manual specialist assessment. Fully automated analysis of retinal OCT from routine clinical practice could provide a promising horizon for diagnosis and prognosis in both research and real-life patient care, following further clinical validation

Funding Apellis Pharmaceuticals.

Lancet Digit Health 2021; 3: e665–75

Published Online September 8, 2021 https://doi.org/10.1016/S2589-7500(21)00134-5

See Comment page e617

*Contributed equally

NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust, UCL Institute of Ophthalmology, London, UK (G Zhang MSc, D J Fu PhD, B Liefers PhD, L Faes MD, S Glinton PhD, S Wagner MD, R Struyven MD, N Pontikos PhD, P A Keane MD, K Balaskas MD); **Department of Ophthalmology, Erasmus University Medical Center, Rotterdam, Netherlands** (B Liefers); **Eye Clinic, Cantonal Hospital of Lucerne, Lucerne, Switzerland** (L Faes)

Correspondence to: Prof Konstantinos Balaskas, NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust, UCL Institute of Ophthalmology, London EC1V 2PD, UK **k.balaskas@nhs.net**

Introduction

Age-related macular degeneration is one of the leading causes of blindness and vision loss globally among people older than 55 years.[1] The non-neovascular subtype comprises nearly 90% of cases of age-related macular degeneration wherein vision loss results from progressive degeneration of the retinal pigment epithelium (RPE) and outer retinal layers. Geographic atrophy is the defining atrophic lesion of advanced non-neovascular age-related macular degeneration.[2] Given its effect on morbidity and absence of proven treatments, geographic atrophy represents an ongoing clinical challenge and disease-modifying therapeutics for geographic atrophy are an intense field of investigation.[3,4] Standardised

Research in context

Evidence before this study

In this study, we set out to develop and validate a deep-learning-based approach for the automatic segmentation of geographic atrophy and its subtypes from optical coherence tomography (OCT) imaging. Such an artificial intelligence segmentation tool would be essential to render real-life monitoring of geographic atrophy progression feasible, as manual segmentation of three-dimensional OCT scans at scale is not possible in clinical practice due to its labour-intensive nature. We did a thorough evidence review on the topic of automatic segmentation of geographic atrophy on OCT. We searched Pubmed, Scopus, Science Direct, and IEEE Explore from database inception to Dec 13, 2020, for studies in English that applied machine and deep learning on OCT imaging for the identification of geographic atrophy. We combined the search terms artificial intelligence ("deep learning" or "machine learning"), OCT ("OCT" or "optical coherence tomography"), segmentation ("segment" or "segmentation"), and geographic atrophy ("geographic atrophy" or "GA"). Studies in languages other than English that included an English abstract were also considered. We found six studies using machine learning to segment geographic atrophy from OCT imaging. We assessed these studies against the following five criteria: sample size, segmentation accuracy performance metrics, ability to segment constituent geographic atrophy features on OCT, ability to inform subtyping as per the Classification of Atrophy Meetings group classification, and performance reporting on external and out-of-sample validation. None of the reported studies fulfilled the latter three criteria and all were based on small sample sizes.

Added value of this study

The current standard for segmentation is manual efforts by human experts, which is time-consuming, labour-intensive, and prone to inter-grader variability. Therefore, machine learning algorithms are being developed to segment individual retinal features that correlate with the presence of geographic atrophy. Although geographic atrophy segmentation algorithms to date have been based on single histological features of geographic atrophy, the model presented here comprises all consensus histological features available in an OCT scan. This fully automated model was trained and evaluated using an external validation dataset assembled from real-world clinical practice and is the largest cohort size available on geographic atrophy to date. In this context, our model shows predictions similar to consensus human specialist graders.

Implications of all the available evidence

Fully automated analysis of retinal OCT images from clinical practice is needed for robust routine assessment of geographic atrophy. Slowing growth rate of geographic atrophy and thus vision loss is an important clinical goal, and the ability to segment geographic atrophy at its early stages could help to maximise the window of therapeutic opportunity. Furthermore, standardised diagnosis and prognosis, facilitated by automatic segmentation, can expedite and enhance both research and clinical care.

assessments of geographic atrophy through ophthalmic imaging are necessary to monitor disease activity and progression in clinical practice, as well as inform outcome measures for clinical trials.

Among existing imaging modalities for geographic atrophy assessment, the most recent reference standard (previously colour fundus photography) for diagnosis and staging of atrophy is spectral domain optical coherence tomography (OCT).[5] This broadly available imaging modality allows cross-sectional morphology of neuroretina, RPE, and choroid to be resolved in three dimensions and acquired in a rapid, comfortable, non-invasive manner.[3] The Classification of Atrophy Meetings (CAM) group—an international consortium of experts in age-related macular degeneration and retinal imaging—have produced OCT-based definitions of geographic atrophy and its subtypes with histological correlates and clinical validation.[5,6] These standardised definitions enable more accurate measurements, monitoring of disease progression,[7] and identification of early biomarkers. For example, incomplete RPE and outer retinal atrophy (RORA) is a distinct subtype of geographic atrophy that precedes complete RORA—the endpoint of atrophy.[5,8] These discoveries have refined disease monitoring and prognosis but, more importantly, enabled identification of earlier stages to serve as opportunities for interventions to prevent vision loss. Atrophy in this context, and consequent vision loss, is irreversible and it is therefore essential to identify patients at high risk for progression to advanced age-related macular degeneration who might benefit from early intervention.

Although the advantages of detailed geographic atrophy characterisation are clear for optimising clinical management, manual grading and spatial delineation of disparate retinal features on three-dimensional OCT volume scans requires specialist training.[9] An OCT scan from a single eye alone typically comprises 49 B-scan cross-sectional images and detailed whole-volume OCT segmentation is laboursome and time-consuming, thus unfeasible in clinical practice. Furthermore, manual segmentation is prone to inter-grader variability, thereby exposing real-world care and clinical trial results to unwanted variation.[10] A promising advance is the use of artificial intelligence to develop models that automatically process OCT images, quantifying each of their anatomical constituents in three dimensions and providing quantitative OCT parameters.[11–15] Automated quantitative OCT generates rapidly accessible, objective data and is thus ideally suited to standardise OCT measurements across institutions.

In this study, we aimed to develop and validate a fully automated deep-learning method for quantitative OCT segmentation of geographic atrophy from OCT,

which independently identifies and quantifies each of the retinal features required for geographic atrophy detection and subtyping as per the CAM consensus definitions. We further aimed to externally validate this method on an independent dataset from real-world clinical practice and compare its predictive performance to human specialist graders, as an indication of clinical utility.

Methods

Study design and dataset

We did a deep-learning model development and external validation study on OCT retinal scans at Moorfields Eye Hospital Reading Centre and Clinical AI Hub (London, UK), in which two datasets from independently recruited patient cohorts with geographic atrophy secondary to age-related macular degeneration were used: one for model development and another for external validation of resultant models. This study was in compliance with the Declaration of Helsinki and reporting guidelines for prediction model and development (TRIPOD; appendix p 17).[16,17]

The automated segmentation models were developed as a post-hoc analysis of prospectively collected OCT images of patients with geographic atrophy enrolled in the international, multicentre FILLY trial (NCT02503332).[18] Briefly, this study is a phase 2, multicentre clinical trial assessing safety, tolerability, and efficacy of intravitreal pegcetacoplan in patients with geographic atrophy secondary to age-related macular degeneration, recruiting patients from 46 sites across Australian, New Zealand, and the USA. Aggregate data for age, gender, and ethnicity of FILLY study participants were obtained from published trial results.[18] A random selection of OCT scans across the trial period (Sept 24, 2015, to July 22, 2016; 984 volumes from 399 eyes of 200 patients including both pre-treatment and post-treatment timepoints; all acquired using Heidelberg Spectralis OCT+HRA; Heidelberg Engineering, Heidelberg, Germany) were split between three graders who had no previous familiarity with the trial dataset; with at least 7 years of continuous experience in OCT image labelling and segmentation following project-specific training and accreditation) at the Reading Centre and Clinical AI Hub of Moorfields Eye Hospital NHS Foundation Trust (London, UK) for manual annotation. For each OCT volume, five B-scans were manually segmented for RPE loss, photoreceptor degeneration, and hypertransmission (the most central B-scan of the entire OCT volume was selected, irrespective of atrophy location, with two B-scans on either side each spaced with five slices; figure 1A). The image dataset was split at patient level: 60% (3024 B-scans from 582 OCT volumes of 120 patients) for training, 20% (958 B-scans, 191 OCT volumes, 40 patients) for tuning, and 20% (1067, 211 OCT volumes, 40 patients) for internal validation. Further details are shown in the appendix (p 2).

Image dataset for external validation

To assess model performance, an external validation dataset was selected from a retrospective cohort of all patients who attended Moorfields Eye Hospital NHS Foundation Trust—a tertiary referral centre with 32 clinic sites serving an urban, mixed socioeconomic and ethnicity population—as part of their routine clinical care and therefore no geographic atrophy treatment was given. 192 OCT volume scans of 192 eyes from 110 patients were taken forward, of which 151 scans were from 151 eyes from 88 patients with geographic atrophy secondary to non-neovascular age-related macular degeneration and the remaining 41 scans of 41 eyes were from 22 patients with macular atrophy associated with neovascular age-related macular degeneration (31 eyes), pigment epithelial detachment (eight cases), and epiretinal membrane (two cases). The inclusion of scans with macular atrophy associated with non-geographic atrophy macular pathologies informed an exploratory evaluation of model performance for OCT biomarkers of macular atrophy associated with other retinal diseases beyond geographic atrophy. As with the model development dataset, five B-scans from each OCT volume were annotated for RPE loss, photoreceptor degeneration, and hypertransmission. Segmentations were done by two expert graders independently, who were masked to patient identifiers and data origin. Further information is provided in the appendix (p 3).

See Online for appendix

Model development

The deep-learning models were implemented as a variation of the U-Net architecture ubiquitously used for medical imaging segmentation (appendix pp 4–5). Briefly, for every pixel of an input image, the model outputs a likelihood estimate for a given feature. Models were trained for each of the morphological features that define geographic atrophy: RPE loss, overlying photoreceptor degeneration, and hypertransmission. A fourth model was trained for direct segmentation of geographic atrophy, in which geographic atrophy was taken to be overlapping regions of RPE loss, photoreceptor degeneration, and hypertransmission (ie, co-occurrence as per A-scan; figure 1B).

Model application

For each two-dimensional B-scan, the output of the model on a pixel level was converted into a one-dimensional binary label, representing the presence or absence of the feature per vertical column (A-scan). The output of the vertical column can be transformed for *en face* overlay based on the B-scan locations (figure 1C). For segmentation of geographic atrophy, two approaches were explored: approach one, in which a model was directly trained to segment geographic atrophy (ie, overlapping regions of manually graded RPE loss, photoreceptor degeneration, and hypertransmission were taken to be geographic atrophy); and approach two, which uses the models for the

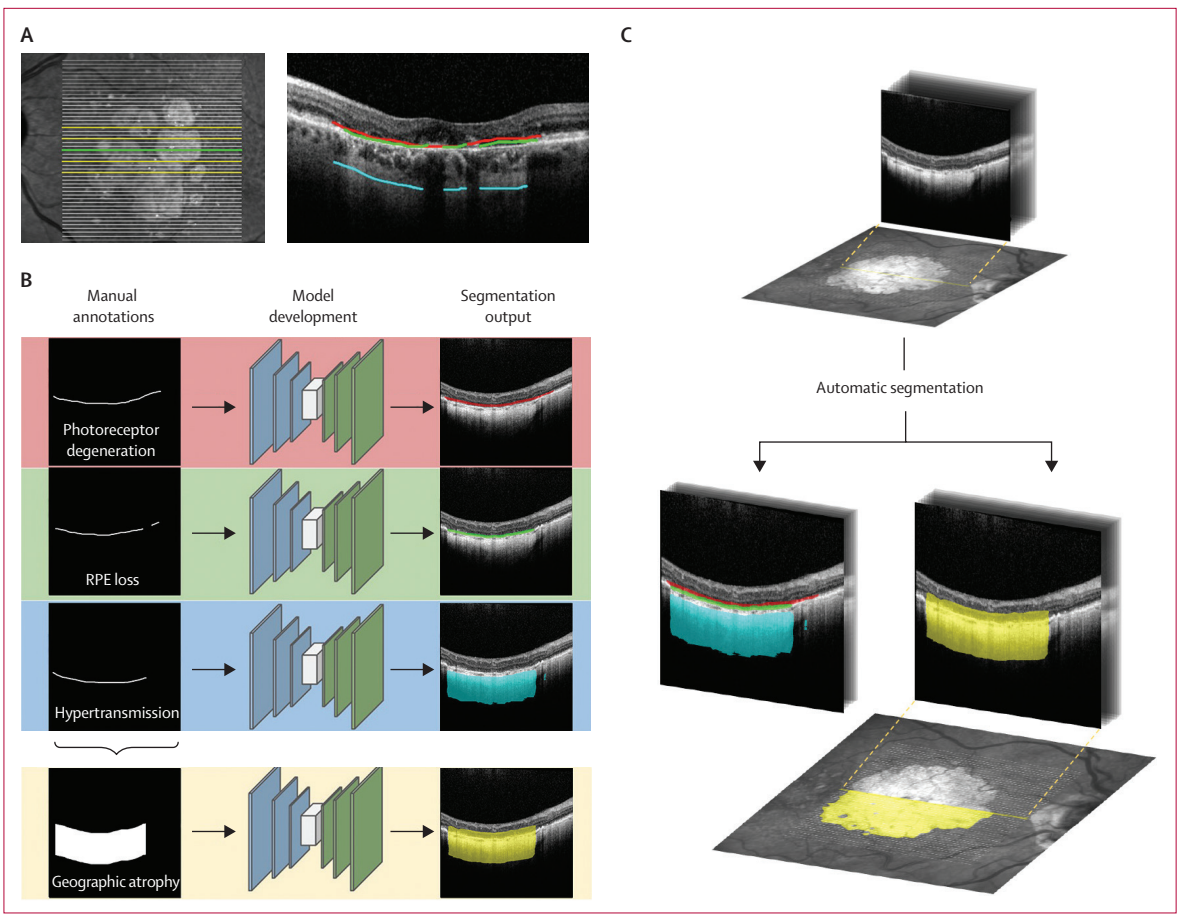

Figure 1: **Development of a deep-learning model for automatic segmentation of geographic atrophy and its constituent retinal features**
The data annotation procedure comprised selection of five B-scans (A, one foveal [green], the most central scan, and two on either side [yellow]), in which the horizontal extent of RPE loss (green), overlying photoreceptor degeneration (red; evidence of outer limiting membrane loss, ellipsoid zone loss, interdigitation zone loss, or outer nuclear layer thinning), and hypertransmission (cyan) were manually demarcated by human graders. (B) Image analysis pipeline. Models were trained on segmentations of the following features: photoreceptor degeneration; RPE loss; hypertransmission; and geographic atrophy, which is taken to be overlapping regions of photoreceptor degeneration, RPE loss, and hypertransmission as per A-scan. (C) For the validation dataset, each of the B-scans of an OCT volume were automatically segmented. Coordinates of the B-scans within a volume permit projection of segmentations onto *en face* fundus photographs. RPE=retinal pigment epithelium. OCT=optical coherence tomography.

constituent retinal features of geographic atrophy (ie, RPE loss, photoreceptor degeneration, and hypertransmission), in which the overlapping of resultant segmentations were taken to be geographic atrophy.

Presence of complete RORA and incomplete RORA were considered at the B-scan level. As per CAM definitions, complete RORA was taken to be any region in which the following were overlapping: (1) RPE loss equal to or greater than 250 µm; (2) hypertransmission equal to or greater than 250 µm; or (3) presence of overlying photoreceptor degeneration. Incomplete RORA was taken to be regions of overlapping RPE loss, photoreceptor degeneration, and hypertransmission but not meeting diameter criteria (1) and (2).

Statistical analysis

Both Dice similarity coefficient (DSC) and intraclass correlation coefficient (ICC) were calculated to measure agreement between model prediction and manual annotations as reference (appendix p 6). The primary outcome was the agreement (DSC and ICC) between model geographic atrophy prediction and consensus of the two independent graders on the external validation dataset. Secondary outcomes comprised agreement between model geographic atrophy prediction and human expert manual annotation on the internal validation dataset; agreement of model predictions for each of the constituent geographic atrophy features on the external validation datasets; mean difference in feature area between predicted and manual segmentations at the OCT volume level; agreement of model predictions for each of the constituent features of macular atrophy associated with other macular pathologies (neovascular age-related macular degeneration, pigment epithelial detachment, epiretinal membrane) on the external validation dataset; categorical detection of features at the

B-scan level (complete RORA, incomplete RORA, or neither; geographic atrophy present and geographic atrophy absent) with performance reported as accuracy (proportion of all predictions that are true positives and true negatives), F1 score:

$$\left(2 * \frac{precision * recall}{precision + recall}\right)$$

and Cohen's κ score. All data analyses were done using R, version 4.1.1.[19]

Role of the funding source

The funder provided input on study design, feedback on the manuscript, and data from the FILLY trial (NCT02503332) used for model development. They did not have a role in data collection for external validation, data analysis, data interpretation, drafting the manuscript, or in the decision to submit the paper for publication.

Results

The dataset for model development comprised 5049 B-scans from 984 OCT volumes taken from 399 eyes of 200 patients undergoing a clinical trial of pegcetacoplan in geographic atrophy secondary to age-related macular degeneration between Sept 24, 2015, and July 22, 2016.[18] Mean age of participants was 79·0 years (SD 7·4) and most were women (106 [53%]; table 1). The model development dataset was randomly split at the patient level. A dataset for external validation (884 B-scans from 192 OCT volumes of 192 eyes from 110 patients) was independently assembled between Jan 1, 2016, and Dec 31, 2019, with data gathered during the clinical care of patients with geographic atrophy secondary to non-neovascular age-related macular degeneration (151 eyes) and macular atrophy associated with other macular pathologies (41 eyes) at Moorfields Eye Hospital. Similar age and gender distributions were observed between the two datasets (mean age 78·3 years [SD 11·1], 58 [53%] women; table 1).

Deep-learning models were developed for each of the constituent OCT features required to define geographic atrophy: RPE loss, overlying photoreceptor degeneration, and hypertransmission. Predictive performances were independently assessed on internal and external validation datasets, in which a wide distribution across all features was observed (appendix pp 7–13).

Correlation coefficients between model predictions and consensus grading were used to evaluate model accuracy. Similar feature performance was observed for each of the models on the external validation set, attaining median DSCs of 0·95 (IQR 0·15) for RPE loss, 0·96 (0·12) for overlying photoreceptor degeneration, and 0·97 (0·07) for hypertransmission (table 2). Promisingly, agreement between model prediction and consensus grading for both internal and external validation datasets were greater than inter-grader agreement (ie, that between two independent human specialist graders: median DSC for RPE loss 0·89 [IQR 0·24], photoreceptor degeneration 0·93 [0·15], and hypertransmission 0·81 [0·28]). Although the models were trained on annotated OCTs from patients with geographic atrophy in the context of non-neovascular age-related macular degeneration, a sub-cohort analysis of other macular pathologies (neovascular age-related macular degeneration, pigment epithelial detachment, and epiretinal membrane) was done and high performance was observed for each feature (table 2). Variability of agreement differences was visualised and evidence of systematic biases were not evident for all features evaluated (appendix p 8). High performance was retained when considering a sub-cohort wherein only one eye and its corresponding volume was included from each patient (appendix p 15) or when comparing against each of the two graders individually (appendix p 16).

Two approaches to segment geographic atrophy were explored in this study. In binary detection of geographic atrophy (ie, present or not) within a B-scan, both approaches had high F1 scores (0·94 [95% CI 0·92–0·96] for approach one; 0·96 [0·96–0·98] for approach two) and accuracy (0·91 [0·89–0·93] for approach one; 0·94 [0·92–0·96] for approach two; figure 2). For geographic atrophy segmentation tasks, predictions of both approaches correlated highly with consensus grading of the external validation dataset. Higher mean DSC and median DSC as well as higher ICC values were observed with approach one (median DSC 0·96 [IQR 0·10]; ICC 0·93) than approach two (0·95 [0·14]; ICC 0·89); yet both outperformed the observed agreement between human graders (0·80 [IQR 0·28]; ICC 0·79; table 3).

	Overall (N=200)	Training (n=120)	Tuning (n=40)	Internal validation (n=40)	External validation (n=110)
Age, years					
Mean	79·0 (7·4)	79·8 (7·3)	77·3 (7·5)	78·2 (7·3)	78·3 (11·1)
Median	80·0 (60·0–97·0)	81·0 (65·0–97·0)	78·5 (60·0–88·0)	79·0 (60·0–90·0)	79·0 (37·0–100)
Gender					
Women	106 (53%)	62 (52%)	24 (60%)	20 (50%)	58 (53%)
Men	94 (47%)	58 (48%)	16 (40%)	20 (50%)	52 (47%)
Ethnicity					
Hispanic or Latino	9 (5%)	3 (3%)	2 (5%)	4 (10%)	..
Not Hispanic or Latino	189 (95%)	116 (97%)	38 (95%)	35 (88%)	..
Southeast Asian	..	..	..	..	4 (4%)
African-Caribbean	..	..	..	..	1 (1%)
White	..	..	..	..	51 (46%)
Unknown	2 (1%)	1 (1%)	0	1 (3%)	54 (49%)

Data are mean (SD), median (range), or n (%). Baseline age of datasets are shown for model development and external validation. The model development cohort featured one eye per patient. Direct comparison of ethnicity was unfeasible as each cohort represented ethnicities with distinct categorical variables.

Table 1: **Baseline characteristics**

	Model *vs* consensus grading						Human grader one *vs* human grader two					
	Number	DSC median	DSC mean	DSC IQR	DSC SD	ICC	Number	DSC median	DSC mean	DSC IQR	DSC SD	ICC
Internal validation dataset												
Photoreceptor degeneration	1042	0·92	0·87	0·12	0·16	0·76	..	..	..	..	..	..
RPE loss	969	0·86	0·76	0·32	0·26	0·68	..	..	..	..	..	..
Hypertransmission	1025	0·91	0·86	0·16	0·16	0·77	..	..	..	..	..	..
External validation dataset												
All												
Photoreceptor degeneration	862	0·96	0·87	0·12	0·21	0·92	877	0·93	0·85	0·15	0·20	0·91
RPE loss	729	0·95	0·87	0·15	0·21	0·86	826	0·89	0·74	0·24	0·31	0·87
Hypertransmission	734	0·97	0·92	0·07	0·15	0·86	838	0·81	0·69	0·28	0·30	0·74
Geographic atrophy only												
Photoreceptor degeneration	680	0·97	0·90	0·18	0·18	0·93	690	0·94	0·87	0·13	0·18	0·91
RPE loss	612	0·96	0·89	0·19	0·19	0·89	668	0·89	0·77	0·20	0·28	0·86
Hypertransmission	615	0·97	0·93	0·14	0·14	0·88	683	0·81	0·71	0·24	0·28	0·71
Other retinal pathologies												
Photoreceptor degeneration	182	0·89	0·78	0·28	0·28	0·89	187	0·88	0·78	0·27	0·24	0·91
RPE loss	117	0·81	0·75	0·25	0·25	0·76	158	0·84	0·62	0·94	0·40	0·90
Hypertransmission	119	0·96	0·88	0·18	0·18	0·80	155	0·76	0·59	0·71	0·38	0·82

Segmentation models were trained for each of the geographic atrophy defining OCT parameters: RPE loss, overlying photoreceptor degeneration, and hypertransmission. DSC and ICC were calculated between resultant models and consensus grading of the internal and external validation dataset. The two human graders on the external validation dataset were also compared. Performance on the external validation dataset was considered as a whole, but also sub-stratified by geographic atrophy secondary to age-related macular degeneration and other pathologies (including neovascular age-related macular degeneration, pigment epithelial detachment, and epiretinal membrane). Number signifies the number of OCT scans used to derive performance metrics. DSC=Dice similarity coefficient. ICC=intraclass correlation coefficient. OCT=optical coherence tomography. RPE=retinal pigment epithelium.

Table 2: **Performance of model segmentation for each of constituent OCT features of geographic atrophy**

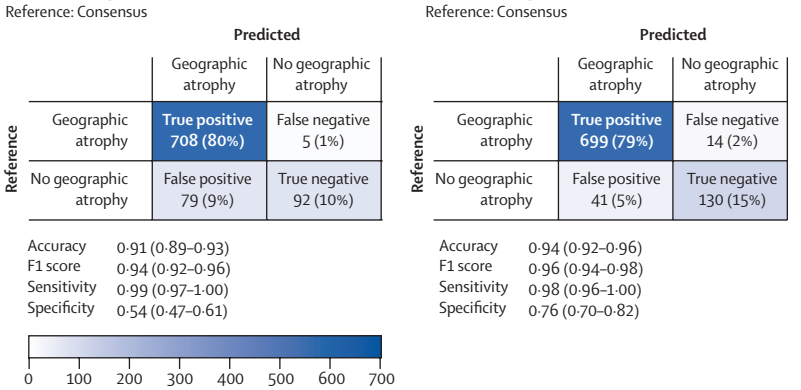

Figure 2: **Automatic segmentation of geographic atrophy**
Confusion matrices for prediction of whether geographic atrophy was present or not in each of the 884 B-scans of the external validation dataset. True reference was taken to be presence of consensus grading. Two geographic atrophy segmentation models were evaluated: approach one trained directly on geographic atrophy segmentation, and approach two wherein segmentation models for RPE loss, photoreceptor degeneration, and hypertransmission were run and overlapping regions were taken to be geographic atrophy. Accuracy, F1 score, sensitivity, and specificity are displayed with 95% CIs. RPE=retinal pigment epithelium.

Consistent with the segmentation models of its constituent retinal features, high performance in detection of overlapping regions (ie, RPE loss, photoreceptor degeneration, and hypertransmission) was retained in OCTs with other retinal pathologies (table 3) and random fluctuation of agreement differences around this mean was observed (appendix p 8).

Automatic segmentation permits quantification of geographic atrophy and thereby differentiation into its subtypes, such as complete RORA and incomplete RORA. For each of the graded B-scans of the external validation dataset, consensus grading found complete RORA to be present in 747 (85%) of 884 B-scans, incomplete RORA in 63 (7%) of 884, and 74 (8%) of 884 to have neither incomplete or complete RORA. With consensus grading as a true reference, both geographic atrophy approach one (accuracy 0·89 [95% CI 0·87–0·91]; F1 score 0·94) and approach two (accuracy 0·87 [0·85–0·89]; F1 score 0·93) had similar predictive performance to that observed between human graders (accuracy 0·88 [0·86–0·90], F1 score 0·93; figure 3).

Automatic segmentation permits rapid detection and quantification of cross-sectional features within entire OCT volumes, which would otherwise be unfeasible with manual segmentation. Time taken for our model to segment geographic atrophy was 0·031 s for a single B-scan and 2·04 s for a 49 slice OCT scan when run on a single graphics processing unit. To illustrate this feature further, entire OCT volumes of four exemplar cases of geographic atrophy secondary to age-related macular degeneration were manually segmented (median time required 43 min per 49 slice volume) and compared with our automatic segmentation models. Areas for each of the retinal features were extrapolated and compared (appendix pp 9, 14). Segmentations from each of the OCTs within a volume can be collectively visualised through projection onto *en face*

	Model *vs* consensus grading						Human grader one *vs* human grader two					
	Number	DSC median	DSC mean	DSC IQR	DSC SD	ICC	Number	DSC median	DSC mean	DSC IQR	DSC SD	ICC
Internal validation dataset												
Geographic atrophy												
Approach one	988	0·84	0·75	0·29	0·24	0·62	..	..	..	..	..	..
Approach two	988	0·83	0·75	0·29	0·25	0·66	..	..	..	..	..	..
External validation dataset												
All												
Approach one	713	0·96	0·91	0·10	0·15	0·93	806	0·80	0·69	0·28	0·30	0·79
Approach two	713	0·95	0·87	0·14	0·20	0·89	..	..	..	..	..	..
Geographic atrophy only												
Approach one	600	0·96	0·92	0·09	0·14	0·94	657	0·75	0·59	0·25	0·38	0·78
Approach two	600	0·96	0·89	0·12	0·18	0·91	..	..	..	..	..	..
Other retinal pathologies												
Approach one	113	0·93	0·86	0·15	0·17	0·89	149	0·80	0·69	0·77	0·30	0·72
Approach two	113	0·85	0·77	0·23	0·25	0·83	..	..	..	..	..	..

For evaluation of segmentation, we compared similarity coefficients (DSC and ICC) between geographic atrophy prediction models and consensus grading of the internal validation dataset and external validation dataset. The two human graders on the external validation dataset were also compared. Performance on the external validation dataset was further stratified by eyes with geographic atrophy secondary to age-related macular degeneration and those with non-geographic atrophy pathologies, including neovascular age-related macular degeneration, pigment epithelial detachment, and epiretinal membrane. Number signifies the number of OCT scans used to derive performance metrics. DSC=Dice similarity coefficient. ICC=intraclass correlation coefficient. OCT=optical coherence tomography.

Table 3: **Automatic segmentation of geographic atrophy**

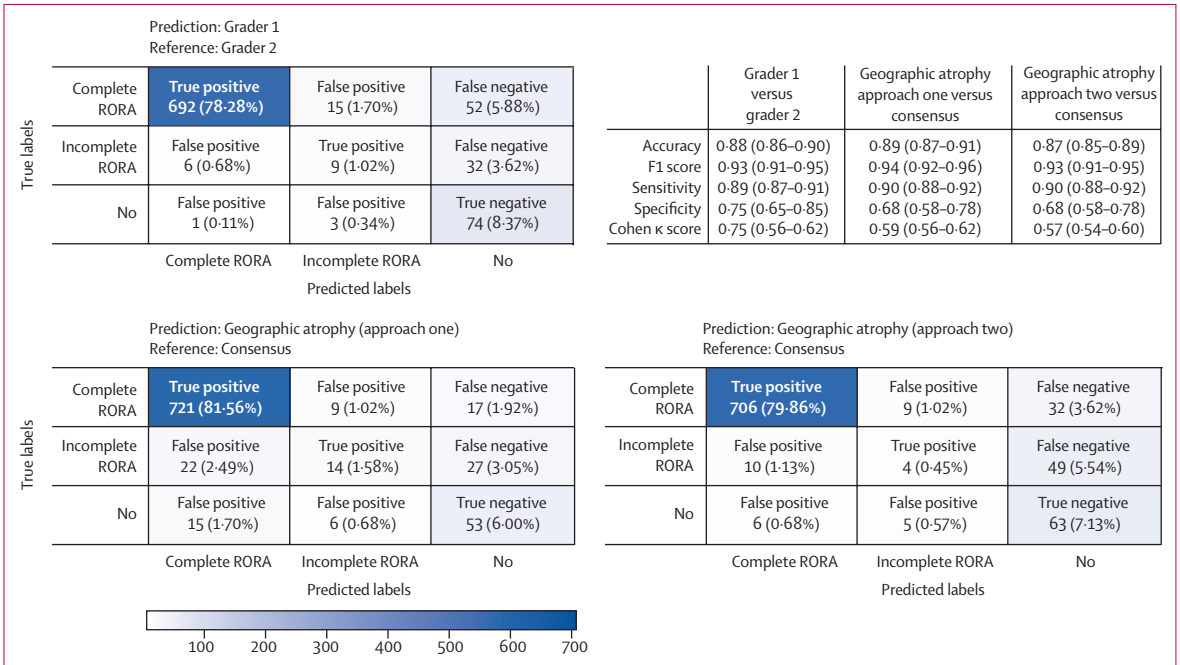

Figure 3: **Confusion matrices for detection of complete and incomplete RORA**

Predictive value in detection of presence of complete RORA, incomplete RORA, or neither were determined for each of the 884 graded B-scans of the external validation dataset. Both geographic atrophy models (approach one and two) were independently assessed with consensus grading as the true reference. Feature detection performance was also assessed for the two human graders, with grader one providing prediction values and grader two as the true reference. RORA=retinal pigment epithelium and outer retinal atrophy.

fundus images. As expected, *en face* projections from manual human grading closely resembled those of the models developed here (figure 4; appendix p 10). Further exploration of automatic segmentation of OCT suggested that this performance would extend to earlier stages of geographic atrophy (appendix p 11), as well as macular atrophy caused by other (non-geographic atrophy) macular pathologies (appendix p 12).

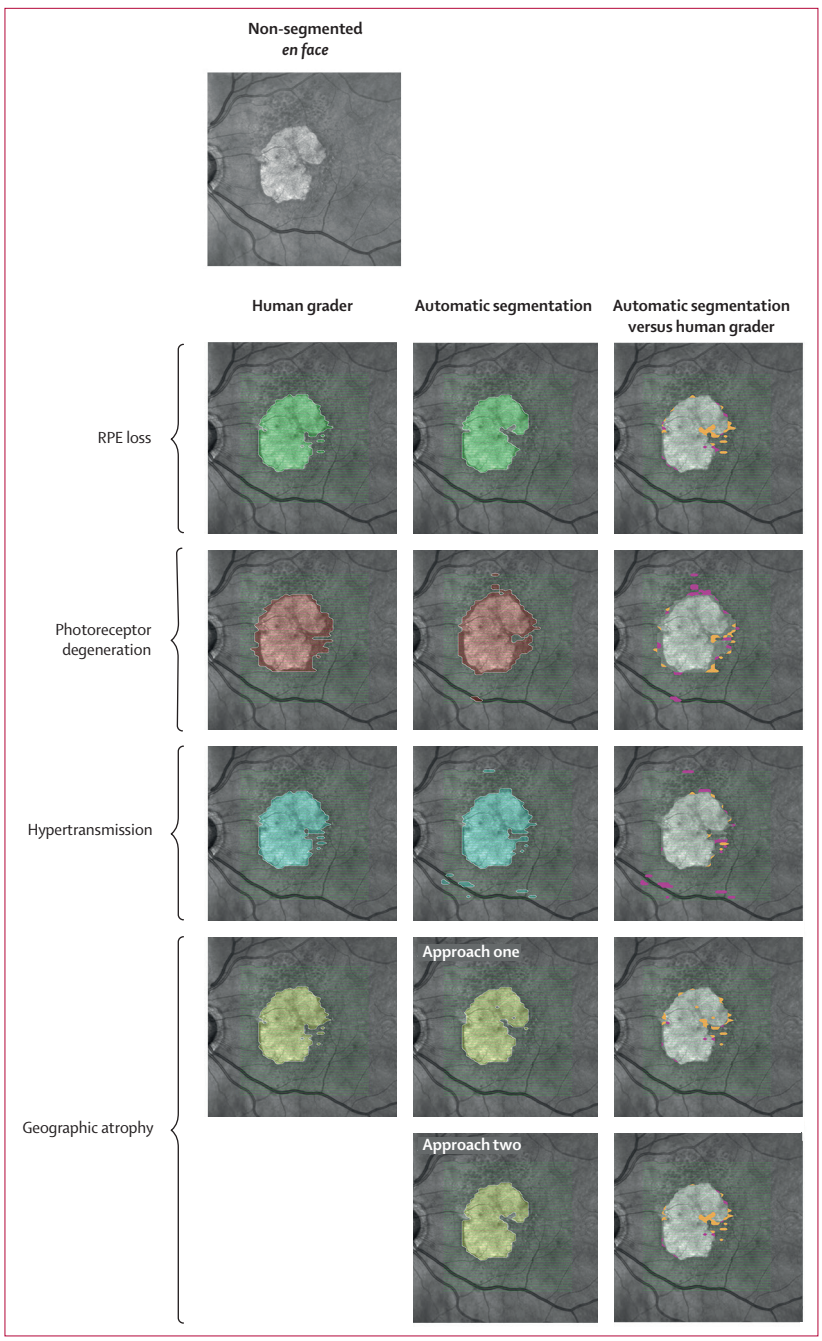

Figure 4: **Automatic segmentation of all images within OCT and projection to *en face***
For four OCT volumes of the external validation set, each of the 49 B-scans were segmented manually by a human grader and automatically with the deep-learning models for RPE loss, overlying photoreceptor degeneration, hypertransmission, and geographic atrophy. Each of these segmentations were projected onto an *en face* fundus photograph (an exemplar depicted) including false positives (pink) and false negatives (orange). OCT=optical coherence tomography. RPE=retinal pigment epithelium.

Discussion

In the present study, we developed a deep-learning tool that automatically processes OCT scans of the retina for detection and quantification of geographic atrophy. To ensure this tool has the sophistication and generalisability required to facilitate real-life clinical management of patients with geographic atrophy and to promote standardisation of clinical trial endpoints for geographic atrophy research, the approach described herein can be scaled across whole-volume OCT B-scans; permits differentiation between geographic atrophy subtypes, as each of the constituent morphological features is considered separately; and is trained and validated on the largest cohort size to date, to our knowledge. Notably, our resultant models show human expert performance on an out-of-sample external validation dataset, indicating potential as a clinical tool to evaluate disease activity with the robustness required for supporting real-life clinical care and assessing clinical trial endpoints.

Development of geographic atrophy is a gradual complex process accompanied by irreversible vision loss. Therefore, initiation of treatment early in the disease process, once therapies become available, is preferable. The CAM consensus group has defined four subtypes of macular atrophy in age-related macular degeneration: complete RORA; incomplete RORA; complete outer retinal atrophy (ORA); and incomplete ORA.[5,8] Definition of the four subtypes of atrophy using OCT imaging requires detection and vertical alignment of the following key features: RPE presence, photoreceptor loss, and hypertransmission. Classification into the four atrophy stages is based on combinations of presence or absence of the three key features and can include measurement thresholds. For instance, complete RORA requires RPE loss and hypertransmission of 250 µm or more, with any overlying photoreceptor degeneration. Complete RORA is considered the endpoint of atrophy and incomplete RORA would be the preceding stage (any of the overlapping features but <250 µm).[8] Areas of incomplete RORA progress and develop into complete RORA over months and years,[20] thus intervention at the incomplete RORA stage could be a window of opportunity to alter the natural course of the disease and prevent further vision loss.[5,8] This potential intervention opportunity depends on the ability to reproducibly identify the constellation of incomplete RORA in OCT scans.

To develop an automated segmentation model that applies the CAM consensus definitions, we established the OCT B-scan as the unit task (ie, the basis for segmenting geographic atrophy and its constituent features) for model training and validation. We then developed individual segmentation algorithms for each of the three constituent morphological features of geographic atrophy and showed high performance for each algorithm on internal and external validation. Furthermore, we showed that geographic atrophy subtypes can be segmented by either direct segmentation of composites (approach one) or differential segmentation with each of the algorithms (approach two). Both displayed highly predictive performance. Although not directly comparable, greater DSCs were observed than

those reported in a systematic review from 2020 of segmentation algorithms for geographic atrophy (range 0·68–0·89).[15]

Currently, slowing growth rate of geographic atrophy and thus vision loss is an accepted clinical trial endpoint. The segmentation approach presented here can support rapid, standardised, and scalable assessment of geographic atrophy over time. Our data suggest that this strategy could be deployed for monitoring progression. Acceptable variation in geographic atrophy area segmentation from manual grading centres in ongoing clinical trials based on two-dimensional imaging modalities is 1·25–2·50 mm².[21–23] In the exemplar cases of whole OCT volume manual segmentation performed here for illustration purposes, model segmentation predictions differ from human graders by mean area differences less than 0·5 mm².

In-depth assessment of geographic atrophy morphology through cross-sectional analysis of the retina, RPE, and choroid via non-invasive OCT has led to the identification of early stage anatomical biomarkers on the B-scan OCT level that predict progression and thus vision loss.[5,24] Such biomarkers are key to informing meaningful clinical trial endpoints for early intervention studies. Manual assessment of OCT scans at this level requires expert graders and is an inherently labour-intensive process associated with intergrader variability and measurement bias. These factors threaten the feasibility of assessing such endpoints in routine clinical practice, or indeed as an efficacy outcome in clinical studies. For instance, the median time required to grade a stack of OCT scans for geographic atrophy at the Moorfields Eye Hospital reading centre is 43 min for a 49 slice volume. With a fully automated three-dimensional quantitative grading system, such as the one presented here, these limitations are overcome. Automatic quantitative OCT generates rapidly accessible, objective data that is ideally suited to standardise OCT measurements across institutions. Such functionality permits interrogation into the interplay between each of the constituent features of geographic atrophy and their relationship with geographic atrophy growth.

Identification and quantification of geographic atrophy constituent features following the CAM consensus guidelines requires depth-resolved, cross-sectional analysis on the B-scan level, which was the unit task used for model development and validation. The resultant model can generate segmentation maps of constituent features of atrophy, allowing clinicians to inspect and visualise an interpretable segmentation. Automatic extraction of three-dimensional quantitative OCT parameters permits geographic atrophy subtyping, which can also be immediately represented and rendered through projection (appendix p 9)—a function that could be readily integrated into existing clinical workflows. Each of the morphological features that define geographic atrophy can be considered separately, thus enabling reliable identification of subtypes. Complete RORA and incomplete RORA were explored in the present study, but there is scope to adapt read-out for identification of the less well characterised, earlier stages of atrophy—complete ORA and incomplete ORA.

There have been previous advancements in semi-automatic and automatic methods for detecting geographic atrophy using different imaging modalities, including two-dimensional colour fundus photographs and fundus autofluorescence, as well as spectral domain OCT.[15] The spectral domain OCT has been adopted by the CAM consortium as the preferred reference standard for diagnosing, characterising, and monitoring progression in geographic atrophy and was the modality used in this work.[25]

The work presented here builds on and advances previous attempts at automatic geographic atrophy detection from spectral domain OCT in three key ways. First, consensus definitions for geographic atrophy based on the widely accepted CAM classification have not previously been used for deep-learning automated segmentation. Most approaches use a proxy (eg, hypertransmission) to segment coarse geographic atrophy regions in the B-scans.[14,26–28] As such, these approaches only serve to identify markers for a generic definition of geographic atrophy and therefore do not have adaptability to differentiate between geographic atrophy subtypes and detect early stages. Secondly, previous sample sizes were limited to 16–56 participants for both training and internal validation;[15] our study included more than double the number of participants than previous studies. Thirdly, this model is the first geographic atrophy clinical tool to evaluate predictive accuracy and robustness beyond the sample used for model development—ie, out-of-sample external validation. In predictive model development, external validation is essential for showing clinical utility.[16] The PROGnosis RESearch Strategy (PROGRESS) group highlights its unique importance stating that "the performance in such a [external] validation study is arguably all that matters, and how a model was derived is of little importance if it performs well".[29] We report the first deep-learning segmentation model to have undergone both internal and external validation on a large real-life clinical dataset and show predictive performance equal to or greater than intergrader agreement of manual segmentation by human expert graders. Following best practice guidelines, the external validation cohort was temporally and geographically distinct from the study population considered in model development.[16,30] As the external validation dataset comprises data collected during clinical care of patients with geographic atrophy, the high predictive performance observed indicates potential for real-world clinical utility. High predictive performance was observed even when considering OCT images featuring macular atrophy associated with other macular pathologies beyond geographic atrophy.

Nonetheless, our study has some limitations. Training and validation of our model was done on OCT scans acquired with the Heidelberg Spectralis, a widely available OCT device globally and the one used to inform the development of the CAM consensus classification. Generalisability of our model to other OCT devices cannot be assumed. The algorithm output is limited to regions captured by the OCT scan and therefore does not detect extra-macular atrophic changes, as might be possible with other imaging modalities such as wide-field infrared and autofluorescence imaging.

This study reports on the first deep-learning segmentation model to apply consensus definitions for the detection, classification, and quantification of geographic atrophy and its constituent features on OCT and to show potential for clinical utility through high performance in a real-life external validation. The validation threshold for clinical application of artificial intelligence decision support systems is determined through an evolving regulatory framework and might require additional evidence from prospective validation and implementation studies, human–computer interaction analyses, and even randomised clinical trials. At present, fundus autofluorescence-derived measurement of geographic atrophy area is the common endpoint in clinical trials for geographic atrophy therapeutics. One would expect output from the RPE loss segmentation algorithm presented here to correlate with fundus hypo-autofluorescence, but this hypothesis is yet to be confirmed. Furthermore, it will be interesting to use emerging alternatives to U-Net segmentation neural network architectures to potentially enhance model performance.

This study presents the development of a novel, fully automated deep-learning method to detect, quantify, and classify geographic atrophy from OCT scans. Predictive performance to a standard similar to clinical experts was reported. Notably, high performance was retained when evaluated on an external validation dataset thereby showing potential for real-life, point-of-care clinical utility. This method could support the management of patients with geographic atrophy and has the potential to facilitate the development of standardised clinical trial endpoints for research into therapy development.

For the **Data Governance framework** see https://www.insight.hdrhub.org

For **PyTorch** see https://github.com/pytorch/pytorch

For **source code** see https://github.com/MoorfieldsInnovationLab/geographic-atrophy-AI

Contributors

GZ was responsible for conceptualisation, data curation, formal analysis, investigation, methodology, validation, visualisation, and review of the manuscript. DJF was responsible for the literature search, study design, data collection, data analysis, data interpretation, data visualisation, and writing of the manuscript. BL was responsible for data collection, methodology, data analysis, and review of the manuscript. LF was responsible for the literature search, figures, study design, data analysis, data interpretation, and review of manuscript. NP was responsible for data analysis, data interpretation, review of the analysis, and review of the manuscript. SG was responsible for data analysis and data interpretation. RS was responsible for data interpretation and review of manuscript. PAK was responsible for initiation of the project, clinical expertise, and review of the manuscript. SW was responsible for data interpretation and review of the manuscript. KB was responsible for conceptualisation, data curation, formal analysis, investigation, methodology, validation, visualisation, and review of the manuscript. All authors had full access to the datasets and DJF, GZ, BL, PAK, and KB verified each dataset during the course of the study. All authors were responsible for the decision to submit for publication.

Declaration of interests

DJF was a consultant for AbbVie, Allergan, and DeepMind. SG is supported by Moorfields Eye Charity (GR001003) and the Wellcome Trust (206619_Z_17_Z). NP is supported by the Moorfields Eye Charity Career Development Award (R190031A) and is the chief executive officer of Phenopolis. SW is supported by an MRC Clinical Research Training Fellowship (MR/T000953/1). PAK is supported by a Moorfields Eye Charity Career Development Award (R190028A) and a UK Research & Innovation Future Leaders Fellowship (MR/T019050/1); receives research support from Apellis; is a consultant for DeepMind, Roche, Novartis, Apellis, and BitFount; is an equity owner in Big Picture Medical; and has received speaker fees from Heidelberg Engineering, Topcon, Allergan, Roche, and Bayer; meeting or travel fees from Novartis and Bayer; and compensation for being on an advisory board from Novartis and Bayer. KB has received speaker fees from Novartis, Bayer, Alimera, Allergan, Roche, and Heidelberg; meeting or travel fees from Novartis and Bayer; compensation for being on an advisory board from Novartis and Bayer; consulting fees from Novartis and Roche; and research support from Apellis, Novartis, and Bayer. All other authors declare no competing interests.

Data sharing

Study data cannot currently be shared publicly due to information governance policies pertaining to real-life clinical data and the absence of informed consent for public clinical trial data sharing. The research team will aim to make the de-identified OCT images available to the scientific community within the Data Governance framework being developed by the National HDR UK INSIGHT Hub. For updates and inquiries please address queries to the corresponding author (k.balaskas@nhs.net). We made use of open-source libraries to conduct our experiments, namely the machine learning framework PyTorch. An end-to-end fully convolutional U-net segmentation model was developed. Model architecture source code is available online. We provide descriptions of the experiments and implementation details in the appendix to allow for independent replication.

References

1 Klein R, Klein BEK, Linton KLP. Prevalence of age-related maculopathy: The Beaver Dam Eye Study. *Ophthalmology* 2020; **127:** S122–32.

2 Gass JDM. Drusen and disciform macular detachment and degeneration. *Arch Ophthalmol* 1973; **90:** 206–17.

3 Holz FG, Sadda SR, Staurenghi G, et al. Imaging protocols in clinical studies in advanced age-related macular degeneration: recommendations from Classification of Atrophy Consensus Meetings. *Ophthalmology* 2017; **124:** 464–78.

4 Chiang A, Lally D. Assessment of progression of geographic atrophy in the FILLY study. *Invest Ophthalmol Vis Sci* 2020; **61:** 4307.

5 Sadda SR, Guymer R, Holz FG, et al. Consensus definition for atrophy associated with age-related macular degeneration on OCT: classification of atrophy report 3. *Ophthalmology* 2018; **125:** 537–48.

6 Göbel AP, Fleckenstein M, Schmitz-Valckenberg S, Brinkmann CK, Holz FG. Imaging geographic atrophy in age-related macular degeneration. *Ophthalmologica* 2011; **226:** 182–90.

7 Cleland SC, Konda SM, Danis RP, et al. Quantification of geographic atrophy using spectral domain OCT in age-related macular degeneration. *Ophthalmol Retina* 2021; **5:** 41–48.

8 Guymer RH, Rosenfeld PJ, Curcio CA, et al. Incomplete retinal pigment epithelial and outer retinal atrophy in age-related macular degeneration: Classification of Atrophy Meeting report 4. *Ophthalmology* 2020; **127:** 394–409.

9 Keenan TD, Clemons TE, Domalpally A, et al. Retinal specialist versus artificial intelligence detection of retinal fluid from optical coherence tomography: Age-Related Eye Disease Study 2: 10-year follow-on study. *Ophthalmology* 2021; **128:** 100–09.

10 Martin DF, Maguire MG, Ying GS, Grunwald JE, Fine SL, Jaffe GJ. Ranibizumab and bevacizumab for neovascular age-related macular degeneration. *N Engl J Med* 2011; **364:** 1897–908.

11 Schlegl T, Waldstein SM, Vogl W-D, Schmidt-Erfurth U, Langs G. Predicting semantic descriptions from medical images with convolutional neural networks. *Inf Process Med Imaging* 2015; **24:** 437–48.

12 De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med* 2018; **24:** 1342–50.

13 Mishra Z, Ganegoda A, Selicha J, Wang Z, Sadda SR, Hu Z. Automated retinal layer segmentation using graph-based algorithm incorporating deep-learning-derived information. *Sci Rep* 2020; **10:** 9541.

14 Hu Z, Medioni GG, Hernandez M, Hariri A, Wu X, Sadda SR. Segmentation of the geographic atrophy in spectral-domain optical coherence tomography and fundus autofluorescence images. *Invest Ophthalmol Vis Sci* 2013; **54:** 8375–83.

15 Arslan J, Samarasinghe G, Benke KK, et al. Artificial intelligence algorithms for analysis of geographic atrophy: a review and evaluation. *Transl Vis Sci Technol* 2020; **9:** 57.

16 Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD statement. *Br J Surg* 2015; **102:** 148–58.

17 Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res* 2016; **18:** e323.

18 Liao DS, Grossi FV, El Mehdi D, et al. Complement C3 inhibitor pegcetacoplan for geographic atrophy secondary to age-related macular degeneration: a randomized phase 2 trial. *Ophthalmology* 2020; **127:** 186–95.

19 Team R. RStudio: integrated development environment for r (computer software manual). Boston, MA: RStudio, 2016.

20 Guymer RH, Markey CM, McAllister IL, Gillies MC. Tolerating subretinal fluid in neovascular age-related macular degeneration treated with ranibizumab using a treat-and-extend regimen: FLUID study 24-month results. *Ophthalmology* 2019; **126:** 723–34.

21 Cheng QE, Gao J, Kim BJ, Ying G-S. Design characteristics of geographic atrophy treatment trials: systematic review of registered trials in ClinicalTrials.gov. *Ophthalmol Retina* 2018; **2:** 518–25.

22 Holz FG, Sadda SR, Busbee B, et al. Efficacy and safety of lampalizumab for geographic atrophy due to age-related macular degeneration: chroma and spectri phase 3 randomized clinical trials. *JAMA Ophthalmol* 2018; **136:** 666–77.

23 Rosenfeld PJ, Dugel PU, Holz FG, et al. Emixustat hydrochloride for geographic atrophy secondary to age-related macular degeneration: a randomized clinical trial. *Ophthalmology* 2018; **125:** 1556–67.

24 Wu Z, Luu CD, Ayton LN, et al. Optical coherence tomography-defined changes preceding the development of drusen-associated atrophy in age-related macular degeneration. *Ophthalmology* 2014; **121:** 2415–22.

25 Csaky K, Ferris F 3rd, Chew EY, Nair P, Cheetham JK, Duncan JL. Report from the NEI/FDA endpoints workshop on age-related macular degeneration and inherited retinal diseases. *Invest Ophthalmol Vis Sci* 2017; **58:** 3456–63.

26 Chen Q, de Sisternes L, Leng T, Zheng L, Kutzscher L, Rubin DL. Semi-automatic geographic atrophy segmentation for SD-OCT images. *Biomed Opt Express* 2013; **4:** 2729–50.

27 Niu S, de Sisternes L, Chen Q, Leng T, Rubin DL. Automated geographic atrophy segmentation for SD-OCT images using region-based C-V model via local similarity factor. *Biomed Opt Express* 2016; **7:** 581–600.

28 Ruiz-Moreno JM, Ruiz-Medrano J, Lugo F, Sirvent B, Flores-Moreno I. Automatic quantification software for geographic atrophy associated with age-related macular degeneration: a validation study. *J Ophthalmol* 2020; **2020:** 8204641.

29 Steyerberg EW, Moons KGM, van der Windt DA, et al. Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013; **10:** e1001381.

30 Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol* 2014; **14:** 40.