nature research

Corresponding author(s): Olaf Ronneberger

□ Initial submission □ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

A sample size requirement of 553 to detect sensitivity and specificity at 0.05 marginal error and 95% confidence was used to inform the number included in the test set. A sample of 997 patients were selected to be part of the gold standard test set against which the human experts and the model were compared.

The total of 15,877 TopCon 3D OCT 2000 scans from 7981 individual patients were eligible for inclusion in the work. An additional 268 Heidelberg Spectralis scans were selected in order to conduct generalisability experiments. The total sample size for training and validation sets was informed by the existing literature and by DeepMind's previous work in the field of machine learning (Mnih et al., 2015; Silver et al., 2016). Today's most powerful deep neural networks can have millions or billions of parameters, so large amounts of data are needed to automatically infer those parameters during learning. Most problems in the medical domain are highly complex as they arise as an interplay of many clinical, demographic, behavioural and environmental factors that are correlated in non-trivial ways. This is even more true for state-of-the art deep learning methodologies that are expected to give the best results (Szegedy et al., 2014).

2. Data exclusions

Describe any data exclusions.

OCT image sets with no diagnostic labels, those containing severe artefacts, or significant reductions in signal strength to the point where retinal interfaces could not be identified were excluded from the present study. Conditions with fewer than ten cases, and data from patients who had manually requested that their data should not be shared, were excluded before research began. For the test set patients who had previously been treated in clinic by the evaluation study participants were excluded from the test set. For more detail please refer to the manuscript methods section.

3. Replication

Describe whether the experimental findings were reliably reproduced.

All 997 patients in the test set for the first device type were randomly selected and were not correlated in any way. The experiments can be interpreted as 997 replicas of a single patient diagnosis. Without retraining the classification network in our framework performance was reproduced on a new test dataset from a second device type of 116 OCT scans. The performance in each case is as follows: Device Type 1 error rate: 55 out of 997 = 5.5%; Device Type 2 error rate: 4 out of 116 = 3.4%.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Samples meeting the inclusion criteria were randomly allocated to training or validation sets. A separate group of patients were randomly selected before creation of the training and validation datasets as an independent test set which was kept separate during model development. Randomisation was on individual patients rather than OCT images: where there were multiple scans for a single patient these were allocated to only one of training, validation or test. For more detail please refer to the manuscript methods section.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Participants in the clinical evaluation of the models were blinded to the ground truth and were not involved in dataset collection; patients who had previously been treated in clinic by the participants were excluded from the test set.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
□	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	A statement indicating how many times each experiment was replicated
□	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
□	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
□	☒	The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
□	☒	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
□	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

The networks used the TensorFlow library with custom extensions (see methods section). Analysis was performed with custom code written in Python.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

The clinical data used for the training, validation and test sets were collected at Moorfields Eye Hospital and transferred to DeepMind data centre in the UK in de-identified format. Data were used with both local and national permissions. They are not publicly available and restrictions apply to their use. The data, or a test subset, may be available from Moorfields Eye Hospital subject to local and national ethical approvals.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

No eukaryotic cell lines were used.

b. Describe the method of cell line authentication used.

No eukaryotic cell lines were used.

c. Report whether the cell lines were tested for mycoplasma contamination.

No eukaryotic cell lines were used.

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

No eukaryotic cell lines were used.

▸ Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	No animals were used in the study.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

Data were selected from a retrospective cohort of all patients attending Moorfields Eye Hospital NHS Foundation Trust, a world renowned tertiary referral centre with multiple clinic sites serving an urban, mixed socioeconomic and ethnicity population centred around London, U.K., between 01/06/2012 and 31/01/2017, who had OCT imaging (Topcon 3D OCT, Topcon, Japan; Spectralis, Heidelberg, Germany) as part of their routine clinical care. For more details please refer to the manuscript methods section.

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