The linear nature of the more-ground-truth effect in explainable deep learning optical coherence tomography image segmentation

Peter M. Maloca[*1,2,3], Maximilian Pfau[1,2,4], Lucas Janeschitz-Kriegl[1,2], Michael Reich[5], Lukas Goerdt[4], Frank G. Holz[4], Philipp L. Müller[3,4,6], Philippe Valmaggia[1,2,3], Katrin Fasler[7], Pearse A. Keane[3], Javier Zarranz-Ventura[8], Sandrine Zweifel[7], Jonas Wiesendanger[9], Pascal Kaiser[9], Tim J. Enz[2], Simon P. Rothenbuehler[2], Pascal W. Hasler[2], Marlene Juedes[10], Christian Freichel[10], Catherine Egan[3], Adnan Tufail[3], Hendrik P. N. Scholl[1,2], Nora Denk[1,2,10]

[1]Institute of Molecular and Clinical Ophthalmology Basel (IOB), Basel, 4031, Switzerland
[2]Department of Ophthalmology, University Hospital Basel, Basel, 4031, Switzerland
[3]Moorfields Eye Hospital NHS Foundation Trust, London, EC1V 2PD, UK
[4]Department of Ophthalmology, University of Bonn, Bonn, 53127, Germany
[5]Eye Center, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, 79106, Germany
[6]Makula Center, Suedblick Eye Centers, Augsburg, 86150, Germany
[7]Department of Ophthalmology, University Hospital Zurich, University of Zurich, Switzerland
[8]Hospital Clínic of Barcelona, University of Barcelona, 08036, Spain
[9]Supercomputing Systems, Zurich, 8005, Switzerland
[10]Pharma Research and Early Development (pRED), Pharmaceutical Sciences (PS), Roche, Innovation Center Basel, Basel, 4070, Switzerland

***Corresponding Author**
Peter M. Maloca,
[1]Institute of Molecular and Clinical Ophthalmology Basel (IOB),
4031, Basel,
Switzerland.
Email: peter.maloca@iob.ch
Tel: +41 61 265 92 14

Abstract

Supervised deep learning algorithms are highly dependent on training data, for which human graders annotate the labels. When annotating digital medical images such as optical coherence tomography images, accurate analysis is crucial to enable correct diagnosis, monitoring, and treatment decisions. These labels, known as ground truth, may be inaccurate and/or ambiguous. This OCT imaging study investigates (1) how size and ambiguity in large ground truth data sets influence the predictive performance of convolutional neural networks (CNNs) and (2) the reproducibility of CNN training. Thirty convolutional neural networks were trained separately with different combinations of ground truths. The Traceable Relevance Explainability (T-REX) technique was used to record and display the results in a way suitable for non-deep learning specialists. The deep learning systems utilized were highly consistent, and their efficiency depended on the ground truth combinations used. Furthermore, a quantifiable linear relationship between ground truth ambiguity and the beneficial effect of having more ground truth (the more-ground-truth effect) was detected.

Introduction

In optical coherence tomography (OCT) imaging[1], a low-coherence laser light is applied to acquire cross-sectional images with micrometer resolution from optical scattering media. Within the medical field, OCT has greatly enhanced research and clinical practice in ophthalmology[2]. Therefore, many diseases that lead to blindness if left untreated, such as glaucoma[3], age-related macular degeneration[4,5], and diabetic retinopathy[6,7], can now be detected earlier, more accurately and in real time. The widespread adoption of OCT in clinical practice has generated huge quantities of OCT data[8], making it the standard for ophthalmic imaging[2]. For example, OCT examinations were observed to increase by a factor of 14 at Moorfields Eye Hospital in London between 2012 and 2016[9,10].

Arguably, the sheer volume of high-resolution OCT digital imaging data predestines ophthalmology to deep learning algorithms, which have been successfully used to deliver valuable clinical information[11,12]. For example, such algorithms have been developed not only to detect specific morphological features of age-related macular degeneration[13] but also to characterize disease activity of the neovascular subphenotypes[14].

In the case of supervised deep learning, the ground truth must be annotated, i.e., contain labels. However, the limited availability of advanced human expert graders for the generation of large ground truth data represents a limitation for the application of deep learning algorithms[15,16]. In addition, we have noticed a certain deep learning

paradox whereby only three to five percent of all available ophthalmological data were used in some crucial deep learning studies[16,17]. Another problem is that the current understanding of what exactly is going on inside the Deep Learning machinery is incomplete which is referred to as the "black box"[18] of deep learning. This poses a hurdle for doctors to implement deep learning into their clinical practice but also for researchers, who would need to better comprehend how a deep learning algorithm achieves a result. The lack of deep learning explainability goes along with mistrust and represents an obstacle to the adoption of artificial intelligence (AI) in the digital medical field. Explainable artificial intelligence (XAI) aims to close this gap by explaining and visualizing individual components of the complex deep learning process[19,20,21]. One element in the toolbox of XAI is the methodology of Traceable Relevance Explainability (T-REX), which was introduced by our group in a previous work[22]. T-REX has already been successfully applied to evaluate the automated segmentation of the ocular compartments including the vitreous, retina, choroid, and sclera ocular compartments[22]. T-REX provides for each individual OCT image a visually readily identifiable tag concerning the annotation performance using Hamming distances among graders and the machine learning algorithm. In particular, T-REX analyzes how uncertainty and ambiguity in the ground truth data affects the training outcomes of artificial neural networks.

This study brings T–REX to the next level of digital medical image analysis by investigating how ambiguity in various ground truth data sets annotated by a higher number of human annotators impacts the predictive performance of convolutional neural networks (CNNs). Thus, we investigated how to apply T-REX to a larger batch of data by means of tracking 30 different CNNs exposed to a broad range of ten

different ambiguous ground truth data sets derived from varying numbers of graders and different ground truth sizes. We wanted to probe whether more data and more graders would induce better deep learning performance and how many graders would be necessary for a particular task.

In addition, we employed T-REX to answer the question of how consistently the performance of a CNN can be reproduced. Hence a reproducible method of deep learning data processing and analysis could help eradicate mistrust in this technology. This point is particularly important, as data on AI reproducibility in the scientific literature are scarce, and many attempts have failed to reproduce a proposed AI method in a second attempt[23]. That is why we decided to reset all parameters after each run for each of the ten CNNs and to repeat the experiments three times independently.

The contributions of our study are as follows. The proposed T-REX methodology could be applied and validated using an extensive number of graders and a variable amount of ground truth data. The deep learning systems utilized were well reproducible, and their efficiency depended on the particular ground truth set used. A previously undiscovered linear relationship between ground truth size, ambiguity in the ground truth, and predictive performance of CNNs trained on this ground truth was detected.

Good predictive performance of the CNNs was observed for the relatively easily recognizable eye compartments, such as the vitreous and the retina. In contrast, the choroid and sclera caused more disagreement, but this could be compensated in part by a higher number of ground truth data.

Hypothetically, a self-optimizing deep learning system in the future could enrich itself and implement an additional learning loop using T-REX to identify problematic samples in the ground truth, and thus possible causes for problems when generalizing to new data. This would enhance the study of learning machines and change their initial learning conditions to reduce further the current limitations of deep learning.

Results

Human graders for ground truth annotation

Out of a total of 10 graders (3 females, 7 males), eight were ophthalmologists, one a veterinary physician, and one a neuroscientist. Their average age was 35.7 years (ranging from 26 to 55 years) and mean work experience with OCT was 10.5 years (ranging from 4 to 24 years). These graders annotated a total of 3200 B-scans (3000 for the ground truth and 200 for the test set), resulting in a total of 3 three segmentation lines per B-scan. This resulted in 3 × 3200 = 9600 line annotations.

Ground truth ambiguity

The obtained data were used to randomly generate ten different ground truth sets (GT sets 1.1, 1.2, 1.3, 2.1, 2.2, 2.3, 3.1, 3.2, 3.3, and 4, referred to as the “ground truth bouquet”). All ten ground truth sets generated for deep learning training contained a similar degree of ambiguity regarding their pixel-wise ground truth labels. The Hamming distance (HD) metric was used to measure the proportion of differently labeled pixels from the B-scans; thus, HD corresponds to one minus the pixel

accuracy between two sets of markers. This became apparent from the mean inter-grader HD calculated from the 200 B-scans of the test set (Table 1, 6th column), where ambiguity ranged from 0.020 (set 1.1 and 1.3) to 0.027 (set 2.2) mean inter-grader HD.

Golden ground truth and learning from ambiguous ground truth

It is quite conceivable that some variation will occur in the annotation of images by human reviewers. Therefore, an absolute unambiguous ground truth can never be fully achieved. We thus proposed for each pixel a majority vote among the human graders, referred to as consensus grade or designated as golden ground truth (GGT)[24]. In this way, each pixel could be assigned to a specific compartment due to the vote of the majority of graders In this context, the three CNNs 4.a, 4.b, and 4.c, which were trained on GT set 4, the largest of the ground truth sets, learned to predict labels that were close to the GGT. The last row of Table 1 shows that the mean HD between these three CNNs' predictions and the GGT (0.011) is smaller than the mean HD between graders and the GGT (0.015) and the mean inter-grader HD (0.023). All calculated HDs refer to the B-scans of the test set.

For CNN 4.a, Fig. 1 shows a heatmap plot visualization of the HDs for each of the 200 B-scans of the test set. The heatmap plot assigns green to small HDs and red to large HDs (Supplementary Fig. S1 shows the same plot with a color-blind-friendly color map). Hamming distances were calculated across all compartments at once (analogous to Table 1 and Fig. 2a). From the color distribution, it is apparent that the HDs between CNN 4.a and the GGT are uniformly green (small HD), in contrast to the HDs between graders and the GGT, which are more variable and in green,

yellow, and orange (larger HD). Heatmap plots for CNNs 4.b and 4.c look similar (plots not shown).

Compartment-specific results

Fig. 2 shows a two-dimensional visualization of the observed HDs on the test set between CNN 4.a, the human graders, and the GGT using multidimensional scaling (MDS) plots. CNN 4.a's predictions (yellow dots) are close to the GGT (blue dots) when considering all compartments simultaneously (Fig. 2a), the vitreous (Fig. 2b), the choroid (Fig. 2d), and the sclera (Fig. 2e). In the case of the retina, CNN 4.a's predictions are not particularly near to the GGT (Fig. 2c). Fig. 2 also shows that mean HDs are one order of magnitude larger in the choroid and the sclera (Fig. 2d, 2e) than in the vitreous and the retina (Fig. 2b, 2c). MDS plots for CNNs 4.b and 4.c look very similar (plots not shown).

Consistency of CNN predictions

The training of artificial neural networks is generally not deterministic because of the randomness associated with the initialization of network weights before training and due to the random order in which training data are presented to a network during training. This study performed three training runs on each of the ten ground truth data sets to investigate the effect of repeated training on the performance of a particular CNN.

Table S1 indicates that the three CNNs trained on each of the ten ground truth sets were relatively consistent in terms of their learned predictions. In average, the difference among the CNN-generated predictions is 0.0062 HD (6.2 of 1000 pixel labeled differently, Table S1 column 7). This variability is ca. 3.7 times smaller than the average variability present among human expert-generated annotations (0.0229 HD or 22.9 of 1000 pixel labeled differently, Table S1 column 3). Fig. 3 shows multidimensional scaling plots based on the data from Table S1. The color map from Petroff[25] is used to provide a better visual accessibility. Each CNN is represented by a triangle, and the triangle's color indicates the ground truth set on which the respective CNN was trained. CNNs that were trained on the same ground truth clearly cluster when considering all compartments (Fig. 3a), the choroid (Fig. 3d), and the sclera (Fig. 3e). Clusters are less visible for the retina (Fig. 3b) and the sclera (Fig. 3c), where HDs are generally smaller.

To facilitate visualization of the reproducibility aspect of this study, we have added supplementary Fig. S2. This illustration shows scatter plots of ground truth size versus (1) the variability among the predictions of CNNs trained on the same ground truth sets and (2) the variability among the human graders' annotations of the respective ground truth sets.

Ground truth size, ambiguity, and CNN predictive performance

Fig. 4 shows the size of a ground truth set plotted against the predictive performance of the CNNs trained on it. The top and bottom of the x-axis indicate ground truth size in terms of the number of B-scans and the number of human graders that contributed

to generating the ground truth set, respectively. The predictive performance (y-axis) is measured as the mean HD between a trained CNN's predictions and the GGT on the test set. The GGT is used as a proxy for the "true" labels and measures the quality of a CNN's predictions. Plots are shown across all compartments (Fig. 4a) and separately for the vitreous (Fig. 4b), retina (Fig. 4c), choroid (Fig. 4d), and sclera (Fig. 4e) compartments.

Two linear regression methods were applied to obtain a more reliable assessment. Each plot includes two linear regression lines. The solid line was calculated based on ordinary least squares (OLS) regression. The dashed line was calculated based on robust Huber regression. The regression lines estimated by the latter may be more reliable, since the data contained some outliers (e.g., green data points in Fig. 4a). The slopes of all regression lines are negative, indicating a possible negative relationship between the number of ground truths and the predictive performance (mean HD between trained CNNs and the GGT). In other words, the more ground truth was used for CNN training, the closer the CNNs' predictions were to the GGT. This effect could be termed the more-ground-truth effect (more-GT effect). Interestingly, this slope (i.e., the benefit provided by additional graders) varies markedly with the eye compartment. The slope is the flattest for the vitreous and the steepest for the choroid.

Fig. 5 shows a plot of the slopes of the regression lines from Fig. 4 versus the ambiguity in the ground truth. The ambiguity in the ground truth is measured by the mean HD among the ten human labelings of the test set, and it is used as a proxy for the true ambiguity in the ground truth. Ambiguity is calculated across all compartments and for each compartment separately (x-axis). Note that the y-axis, which plots the more-GT effect, is inverted. This is because large negative values on

the y-axis indicate stronger effects. Slopes from OLS and Huber regression from Fig. 4 are plotted as blue and yellow dots, respectively, in Fig. 5. Two linear regression lines fit through the blue and yellow dots (using OLS regression). It is apparent that the compartments with higher ambiguity in the ground truth are associated with a steeper more-GT effect (i.e., CNN will make better predictions when trained on more ground truth).

Note that Fig. 4 illustrates the more-GT effect; that is, how having more ground truth for CNN training affects the CNN's predictive performance. The linear regression lines in Fig. 4 aim to illustrate this effect by assuming a linear relationship. In contrast, Fig. 5 illustrates how the more-GT effect relates to the ambiguity in the ground truth and that this relationship is a linear one (shown by linear regression lines in Fig. 5).

Discussion

Various studies have shown that deep learning is ideally suited to segment and classify OCT data at a level at least similar to that of human examiners[26,27]. Despite its enormous scientific success with more than 20,000 medical publications, deep learning has hardly been used in routine clinical practice to date[28,29]. For physicians and practitioners to trust deep learning application, they need to understand how deep learning makes predictions. Moreover, only about eight percent of the publications reported on the reproducibility[30] of the deep learning method used.

In spite of all the deep learning progress, the "black box" nature of deep learning is posing an obstacle to the widespread implementation of deep learning in medical applications, consequently also for the use in machine learning analysis of OCT

images. This is where XAI[31,32] comes into play. XAI aims at explaining the decisions AI applications make in a way that is comprehensible to humans. One aspect of explainability is how deep learning deals with uncertainty and ambiguity in ground truth data[22]. This is an omnipresent deep learning phenomenon since most ground truth data sets exhibit some level of uncertainty owing to various reasons like the quality of the data to be annotated, the difficulty of the annotation task, or human error. To partly open the deep learning "black box," we proposed a quantitative data analytics methodology, marked as T-REX[22].

Hence, to promote the further use of artificial intelligence in health-care, reproducible and traceable models could increase trust in these deep learning systems[33]. This would, in turn, lead to trustworthy AI[13] and an advance of digital medicine in general.

Learning behaviour, ground truth size and ground truth ambiguity

Since the mean HD between the predictions of the CNNs used and the annotations of the human graders is smaller than the mean inter-grader HD, it appears that the CNNs learned some kind of average among the human graders' annotations. This could be a general trend. When considering all 30 CNNs that were trained in this study, 23 showed a mean HD compared to human graders that was smaller than the corresponding inter-grader HD (Table S1). This means that compared to the human graders, these CNNs exhibited a distinct averaging behavior, which was also observed in our previous study[22]. Thus, the CNNs positioned themselves, in some sense, *in between* the human graders with regard to the predictions they made.

For unambiguous ground truth, such as the internal limiting membrane (ILM) border, more graders and, consequently, more ground truth data appeared to improve the predictive performance of the CNN. With enough effort, human graders can

determine the ILM almost pixel-perfectly. It is, however, very time-consuming to draw the ILM perfectly, so the graders are often a few pixels above or beneath the actual line. Therefore, we used the GGT as a proxy for the actual line. With more data, the CNN was able to learn the exact location of the ILM better even if the ground truth contained some noise. The errors made by the graders were errors that could be identified as such. In this case, the ground truth was not ambiguous but just faulty. Nevertheless, it was interesting to observe that the CNN was obviously able to learn the true objective despite the deviating annotations[34]. These findings are consistent with the results of a previous study[22] that also examined the influence of ambiguity in ground truth labels on the predictions learnt by a CNN, albeit on a smaller scale.

The situation appeared to be different for ambiguous ground truth, such as with regard to the choroid-sclera interface (CSI). We again used the GGT as a proxy for the true ground truth line. This time, it was impossible for human experts, based on the OCT imagery alone, to determine the “true” border. However, the CNNs were still able to learn to predict closer to the GGT with more annotated data (Fig. 4), and they learned better with more data for all four compartments. This can be seen from the negative slopes of all regression lines in Fig. 4. The steepness of the slopes of the regression lines indicates the effect of having more ground truth. The effect is strongest in the choroid, which has the highest ambiguity. The slopes of the linear regression lines indicate the strength of the more-GT effect. Interestingly, this effect appears to be linearly correlated to the ambiguity in the ground truth (Fig. 5). The more ambiguous the ground truth is, the more it helps a CNN to have more of it to learn better predictions.

The relationship between the size of a ground truth data set and the predictive performance of a CNN (Fig. 4) is arguably not a linear one. At some moment, the predictive performance of a CNN will converge with more ground truth. However, the linear approximation in Fig. 4 reveals an interesting link between ground truth size, CNN predictive performance, and ground truth ambiguity. In a further step, this relationship could be investigated outside the “sweet spot” investigated in this study, such as with much smaller or bigger ground truth sets. This could also reveal if there is an irreducible error in settings where ambiguous ground truth data are used for CNN training.

Investigating the interaction between ground truth size, ground truth ambiguity, and predictive performance of machine learning algorithms also has applications outside the field of ophthalmology. In natural language processing, for example, supervised machine learning is often applied in automated language translation (e.g., text-to-text[35], speech-to-text[36], and speech-to-speech[37]). Having diverse ground truth sets annotated or labeled by numerous people is even desired in these fields, since it captures more of the diversity with which a language is used. T-REX can be leveraged in these cases to shed light on how machine learning algorithms are affected by the number and diversity of human annotators and the size and ambiguity of the resulting ground truth sets.

Reproducibility respectively consistency of CNN training

An interesting aspect of deep learning algorithms appears to be their ability to perform a task comparable to humans even without explicit coding[18]. This almost "automatic execution by a machine" can suggest a rigid and error-free deep learning sequence that in part may not exist. Therefore, it was important for us to investigate how deep learning algorithms learn and whether their result is repeatable. Although deep learning

is full of randomization by means of random initialization, data augmentation, and random noise introduction, the experiments performed in this study showed a high agreement. In addition, to obtain more confidence, we repeated each experiment for each of the ten CNNs not twice but three times, although we did not encounter this approach in the current deep learning literature[30]. The obtained results were robust and can be trusted[23].

Because there is no current consensus regarding reproducibility[30] as a term in machine learning[38], we follow the recommendations of the National Academies of Science, Engineering, and Medicine[39]. Therefore, we would rather propose the term “consistency of deep learning training,” since the obtained results were similar but not identical. This deviation, even if only to a small extent, indicates that the CNN training cannot be completely "reproducible". We show this subtle nuance, for example, in Fig 3, where a small yet measurable variation between the runs is detected, despite identical training ground truth data for each run. Hence, in the best case, a deep learning algorithm shows a certain degree of “deep learning consistency”, rather than reproducibility, in terms of its training process. More specifically, given a ground truth set, the average HD among the CNNs trained on that set was 0.006 (mean of 7th column of Table S1); thus, training two CNNs on the same ground truth sets leads to predictions that differ in six out of 1000 pixels. On the one hand, this shows some level of consistency, in comparison to the annotations of two human graders, which differed on average in 23 of 1000 pixels (mean of the 3rd column of Table S1). The deep learning variability is 3.5 times smaller than the average variability present among human expert-generated annotations. See also Fig. S2. On the other hand, this result shows that CNN training is not fully reproducible, which to our knowledge has not yet been studied systematically with regard to OCT data analysis. However, the

inconsistencies of the predictions made by those CNNs are well within the range observed in human-generated annotations, which is a good achievement in itself. This is also well illustrated in the MDS plots (Fig. 3). The three training runs on the same dataset are mostly grouped close to each other. The variations mentioned may not be relevant to the clinical use of deep learning, but we add a novel element to better understand variations in deep learning performance.

In conclusion, the repeated training process of an artificial neural network for semantic image segmentation of OCT images shows a relatively stable convergence to an optimum with respect to the predictions made on the test set. This study demonstrated that deviations could be detected depending on the task's difficulty, but the overall variations were relatively small (e.g., for the vitreous). Thus, a training consistency could be observed, as the CNNs always converged to a certain value. Nevertheless, no CNN was able to reach the GGT completely. A further step could be to investigate the convergence with respect to artificial neural network weights.

In machine learning and statistics, it is best practice to investigate how model fitting/training is affected by uncertainty arising from the random components of the fitting/training process[40,41]. However, in deep learning these effects are rarely considered or studied systematically. This could even lead to a relatively strong publication bias since it's tempting to publish just the model with the best performance neglecting training attempts that did not yield quite as good outcomes. Regarding the importance, significance, and potential of deep learning in medical applications, we suggest including a systematic investigation of the repeatability/consistency of deep learning models as a standard procedure in future studies. The proposed T-REX approach could be beneficial in this regard, because it

allows for a quantitative evaluation of the repeatability/consistency of deep learning experiments. This repetition of deep learning experiments might be time consuming, but it could help define new deep learning standards and prevent methodological pitfalls[38]. However, we suggest that in the future a similar measurement of the repeatability of a deep learning method should be part of an evaluation process to confirm its reliability.

There are limits to the current work. In some instances, the intense pigmentation of the choroid made it difficult to delineate reliably between the choroid and the sclera. This can lead to erroneously large or small compartments and influence the results. Nevertheless, the match between the graders was acceptable. Technically, it is currently not possible to precisely align the location of a medical image annotation with its true anatomical location without biopsy and to injure it in such a way. Another weakness is that the graders did not undergo a consensus finding to determine the exact boundaries mentioned. Such a consensus finding would probably have led to better agreement between the graders and consolidated the results. However, the goal of this study was not consolidation; it was to investigate uncertainties and how a deep learning system responds to them. Furthermore, only relatively experienced graders were involved. Another limitation could be induced by the random selection of the individual B-scans. The B-scans were labelled independently, and multiple B-scans could be present from one eye. Since a volume may contain important contextual information compared to an isolated B-scan, inaccuracies are possible. This should also be the case for volumes without pathology - for example, the outer border of the choroid may be easier to delineate in some scans than others.

In summary, the proposed T-REX methodology was successfully applied to a large dataset, and it was not only shown, but also quantified how more data and more graders induce a better deep learning learning-performance. It appears from the smaller mean HD between the predictions of CNNs and the human graders' annotations that the CNNs learned some average of the human graders' annotations. Also, it was found that the performance of a particular CNN can be reproduced when using the identical training data repeatedly. A novel and interesting quantitative relationship was identified whereby more uncertainty per grader requires more data from more graders for similar results. This relationship appears to be linear; thus, the effect can be referred to as the linear more-ground-truth effect in deep learning OCT image segmentation. T-REX, in general, is a data analytic methodology that enables the quantitative study of (1) ambiguity in the ground truth data and (2) deep learning repeatability/consistency. As such, T-REX is applicable to the medical domain but also to areas outside the medical field.

Methods

Animals and husbandry

Data were retrospectively collected from 44 healthy and untreated cynomolgus macaques (17 females, 27 males) of Mauritian genetic background with an age range of 30–50 months and weighing between 2.5 kg and 5.5 kg. All animal investigations were conducted in strict compliance with the applicable guidelines of the US National Research Council and the Canadian Council on Animal Care Studies. Only OCT image data from treatment-naïve cynomolgus monkeys of both sexes were retrospectively reviewed for the purpose and application in the current study. No additional animal experiments were required for the current study. Primary experiments were reviewed and approved by the Institutional Animal Care and Use

Committee (IACUC) of each institution, namely Charles River Laboratories Montreal, ULC Institutional Animal Care and Use Committee (CR-MTL IACUC), IACUC Charles River Laboratories Reno (OLAW Assurance No. D16-00594), and the IACUC (Covance Laboratories Inc., Madison, WI, USA; OLAW Assurance #D16-00137 A3218-01). Animals were treated and handled in rigorous compliance with the guidelines of the US National Research Council and the Canadian Council on Animal Care. Animals were accommodated in groups in stainless steel cages that met the European housing standards described in Annex III of Directive 2010/63/EU. Room temperature was kept constant between 20 °C and 26 °C; humidity was between 20% and 70%, and the light–dark cycle was a standard 12:12 hours. Animals were fed a standard diet of pellets supplemented with fresh fruits and vegetables. Tap water was freely offered through an automatic drinking system after being treated by reverse osmosis and ultraviolet irradiation. Psychological and environmental enrichment was provided to the animals, except during study procedures and activities.

OCT imaging data

Only macula OCT scans were extracted from the existing data library, acquired with Spectralis HRA+OCT- (Heidelberg Engineering, Heidelberg, Germany) with scan angle 20 degrees, 25 raster-line B-scans, scan length 5.3 mm, scan depth 1.9 mm, 512 pixels × 496 pixels, and activated automatic real-time tracking (ART) averaged for 30 scans. OCT data from healthy animals were included that showed complete visualization of the macula so that all four compartments (vitreous, retina, choroid, and sclera) were depicted. The image quality according to the manufacturer's software was required to be a minimum of 25. OCT data were excluded in the presence of any retinal or choroidal pathology that could be visualized with OCT.

Human expert grading of OCT images

Ten graders from five different ophthalmic centers (University Basel, Switzerland, Roche Basel, Switzerland, University Bonn, Germany, Moorfields Eye Hospital London, UK, and University Zurich, Switzerland) contributed to the OCT image annotations. Each of the ten human graders independently labelled B-scans out of 12 OCT volumes derived from healthy eyes using a custom written and password-protected online annotation tool. Graders 2 and 3 were a veterinarian and a biologist, respectively, who worked daily with OCT in preclinical research. All other human graders were trained retina specialists (ophthalmologists). The labels generated by graders 1, 2, and 3 had already been used in a previous study[22].

Since an OCT volume consisted of 25 B-scans, each grader labelled a total of 300 B-scans. Therefore, a total of 10 × 25 × 12 = 3000 B-scans were labelled by the human expert graders. For each grader, their annotated data were split into a training set and a validation set that were later used for CNN training. The training set of each grader consisted of nine OCT volumes with a total of 225 images, and the validation set consisted of three OCT volumes with a total of 75 B-scans.

Additionally, all graders labelled the same test set consisting of a total of 200 B-scans from eight additional OCT volumes.

Ground truth generation

This study investigated how the predictive performance of a CNN is affected by the size of the ground truth set it was trained on and the number of human graders that contributed to that ground truth. For this purpose, ten different ground truth sets were created with a varying number of contributing graders and B-scans (Table 1).

[Insert Table 1 about here]

Each ground truth set was created by randomly selecting the required number of human graders (random sampling without replacement, Table 1, column 5). In that way, two constraints were satisfied. First, the ground truth datasets 1.1, 1.2, and 1.3 were mutually exclusive (meaning a grader contributed at most to one of these sets). Second, each grader contributed ground truth to at least four and at most six out of the ten ground truth sets to ensure a balanced experimental design. The annotation process of the ground truth by the human graders was described in detail in a previous work[22]. In short, the human graders annotated each B-scan separately using a password-protected web browser-based graphical user interface. Each B-scan showed the posterior eye pole with the four compartments comprising the vitreous, retina, choroid, and sclera. Annotating means pixel-wise drawing of three lines on each B-scan with a computer mouse, as follows: (1) the ILM between vitreous and retina, (2) the hyporeflective choriocapillaris (CCi) between retina and choroid, and (3) the choroid-sclera interface (CCI) between choroid and sclera. Subsequently, these annotations were turned into pixel-wise label maps. B-scans and label maps were reshaped to a spatial resolution of 512 × 512 pixels, which corresponds to the required input format of the CNNs.

Definition of golden ground truth

The human graders generally annotated the B-scans of the test set slightly differently. This is because (1) the labeling task is generally difficult (drawing lines on images with more than 100,000 pixels) and (2) each grader has a different professional background that acts in some sense as prior knowledge of how to draw the lines. This means the annotations generated by the human graders can be noisy and/or ambiguous. An unambiguous truth for the position of the lines (the labels) may

exist, since the borders of the four compartments are physiologically and anatomically well-defined. However, it is generally impossible to locate them unambiguously on a B-scan without dissecting an eye. Thus, in this setting, a ground truth that is unambiguously true cannot be obtained. The best that can be done to approximate the “true” labels is possibly to use a majority vote among the human graders, assuming that, on average, a human grader’s annotations are unbiased. This work thus defined the GGT for the labels of the test set of 200 B-scans as the majority vote among the annotations of the ten human graders on a per-pixel basis.

Deep learning

Each of the ten ground truth sets was used to train three CNNs; thus, a total of 30 CNNs were trained in this study. All CNNs used the same neural network architecture and similar learning settings. The architecture of the CNNs, which was basically a U-Net[42] architecture extended by an additional max-pooling layer to account for the large spatial input, was the same as in a previous study[22]. For all CNNs, a mini-batch size of eight and a learning rate of 6×10^{-5} was used. Training was done on an NVIDIA GeForce GTX TITAN X GPU using tensorflow v1.14 with python 3.5. For more details about the neural network architecture and learning settings, refer to the previous study[22]. Each of the ten ground truth sets used in this study fell into one of four size categories (675, 1125, 1575, or 2250 B-scans) in the training set. To ensure that each B-scan in the training set was seen approximately the same number of times, CNNs trained on bigger ground truth sets were trained for longer than CNNs trained on smaller ground truth sets. In each training, each B-scan was seen approximately 23.5 times (see Table 2). For each of the 30 trained CNNs, when training was stopped, the loss had already reached a plateau. CNN training is usually non-deterministic. This is mainly because of the following two factors: a random

initialization of weights before training and a random order of mini-batches during learning. In this study, each CNN arrived at a different (local) optimum when training was finished.

[Insert Table 2 about here]

T-REX methodology

This work used the T-REX methodology[22] to investigate how ground truth size and ground truth ambiguity impact CNNs' predictive performance. T-REX measures the difference between two sets of pixel-wise labels by the HD and visualizes this difference by means of MDS and heatmap plots.

The HD, in short, is a metric that measures the proportion of elements that are labeled differently between two sets of labels. For example, if two pixel-wise labelings of an image with a resolution of 10 × 10 pixels differ in a single pixel, the HD between the two sets of labels would be 0.01. HDs can be calculated across multiple label categories at once or for a single label category versus the others. In the latter case, a pixel-wise label map is turned into a binary map where a pixel either belongs to the respective category or not. Mean HDs were calculated between pairs of test set labeling from the human graders, the GGT, and the CNNs' predictions. The HD is a metric by mathematical definition and therefore lends itself to visualizations by means of metric MDS.

MDS is a method to visualize the distances among data points by projecting them into a two-dimensional, Euclidian coordinate system while preserving the distance among the data points as well as possible. In this study, individual data points comprised the labeling of the test set. Another way to visualize HDs is through heatmap plots. A HD heatmap plot basically visualizes the HD among two or more

labelings by using a heatmap. For a detailed description of T-REX, including HD, MDS, and heatmap plots, refer to the previous study[22].

Reproducibility and consistency after three training runs

When training the CNNs, there are random elements; that is, the weights are initialized randomly and the mini-batch ordering is also random. Therefore, two training runs with the same CNN architecture, dataset, and number of training epochs will always yield slightly different results. Machine learning models are often retrained when new data are available, but in health-care applications, it can be undesirable for the models to change their predictions over different training runs. Even if the ground truth data were fixed, reproducibility could still be challenging. In this study, we wanted to examine how similar the CNNs' predictions are when they are trained on the same datasets and for the same number of epochs. The ten different datasets were used to train repeatably the same CNN. We initiated new training with the CNN for three training runs for each dataset and deleted all parameters after each run. A training run was repeated if it did not converge, it overfitted, or it yielded otherwise unsatisfactory results. To compare the training runs, we let the trained CNNs predict the labels of the test set to calculate the different HDs from those labels for each run. Thus, for each training run, we calculated the mean HD between the graders of the dataset, the mean HD between graders and GGT, the mean HD between the trained CNN and the graders, the mean HD between the trained CNN and the GGT, and the mean HDs between the different training runs on the same ground truth dataset.

Ground truth size, ambiguity, and CNN predictive performance

This study investigated how the size of a ground truth data set affects a CNN's predictive performance. For this purpose, the predictive performance was measured by calculating the HD between the labels predicted by each of 30 CNNs to the labels of the GGT on the test set (the smaller the HD, the better the prediction). The 30 HDs obtained were plotted against the size of the respective ground truth sets (and the number of contributing graders) in scatter plots.

Linear regressions were performed to investigate the relationship between ground truth size and CNN predictive performance. In addition to the Ordinary Least Squares loss function, the robust Huber loss function[43] was used to fit linear regression models because the data had some outliers. Both Huber and OLS regression were fit for all the compartments together and individually for the vitreous, retina, choroid, and sclera compartments, whereby steeper linear regression lines indicated stronger effects of more ground truth. The effect of the ground truth size on predictive performance could be referred to as the more-GT effect.

The relationship between ambiguity in the ground truth and the more-GT effect was investigated by plotting ground truth ambiguity versus the slope of the regression lines (the more-GT effect) in scatterplots. Ground truth ambiguity was measured as mean HD among the ten annotations of the test set labels generated by the ten human graders. Moreover, linear regression lines were fit using OLS loss.

Tables

Table 1. Summary table of the ten ground truth data sets used in this study.

GT set	Number of human graders	Size train set	Size val. set	Contributing graders	Mean HD inter-grader	Mean HD grader-to-golden GT	Mean HD CNN-to-grader	HD CNN-to-golden GT
1.1	3	675	225	1, 2, 3	0.020	0.016	0.015	0.013
1.2	3	675	225	4, 5, 9	0.024	0.014	0.020	0.014
1.3	3	675	225	6, 7, 8	0.020	0.017	0.029	0.038
2.1	5	1125	375	1, 3, 6, 8, 10	0.023	0.014	0.018	0.012
2.2	5	1125	375	1, 2, 3, 7, 9	0.027	0.017	0.024	0.020
2.3	5	1125	375	2, 4, 5, 6, 9	0.023	0.015	0.018	0.010
3.1	7	1575	525	1, 4, 5, 7, 8, 9, 10	0.025	0.016	0.033	0.030
3.2	7	1575	525	2, 3, 4, 5, 6, 7, 10	0.022	0.014	0.020	0.014
3.3	7	1575	525	1, 2, 3, 6, 8, 9, 10	0.023	0.014	0.022	0.016
4	10	2250	750	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	0.023	0.015	0.019	0.011

Notes: On each ground truth (GT) set, three convolutional neural networks (CNNs) were trained. The first column indicates the GT set names. The second column indicates the number of human graders who contributed to each GT set's labels. The third and fourth columns indicate training and validation set sizes, respectively. The fifth column indicates which graders contributed to each GT set (numbered from 1 to 10). The sixth column indicates mean Hamming distances (HDs) among the labelings of the contributing graders. The seventh column indicates the mean HD between the labelings of the contributing graders and the golden GT. The eighth column indicates mean HD between the labelings of the contributing graders and those of the three CNNs that were trained on the respective GT. The ninth column indicates mean HD between the three CNNs trained on the respective GT and the golden GT. All HDs were calculated on the test set of 200 B-scans.

Table 2. Summary table of the number of training steps and epochs used in convolutional neural network training on the different ground truth data sets.

Training set	Number of human graders	Number of B-scans (training set)	Number of B-scans (validation set)	Training steps	Epochs
1 ×	3	675	225	2000	23.7
2 ×	5	1125	375	3300	23.5
3 ×	7	1575	525	4600	23.4
4	10	2250	750	6600	23.5

Figure captions

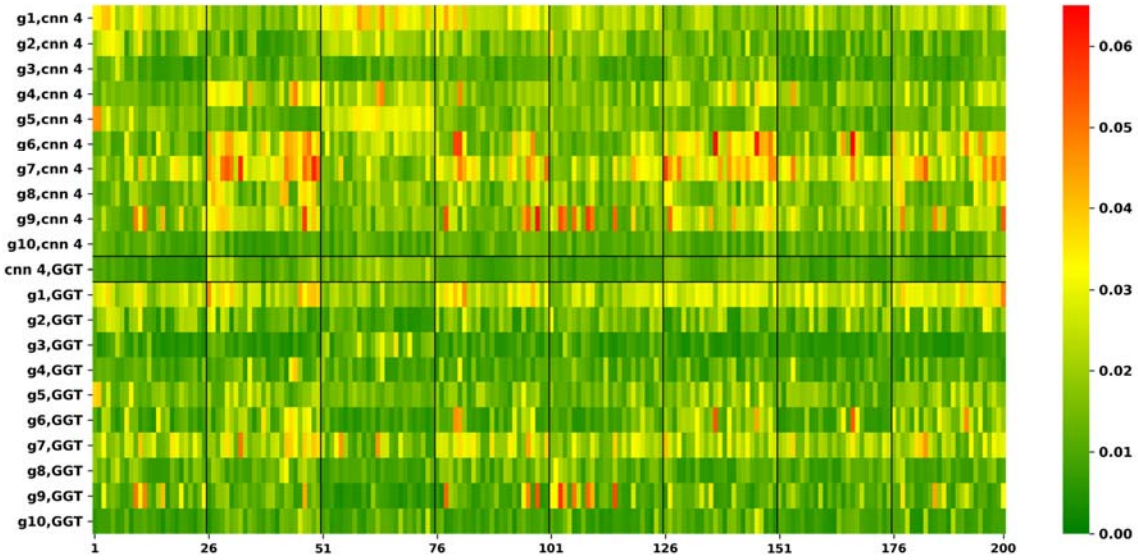


Fig. 1. Heatmap plot visualizing the mean Hamming distance (HD) for each of the 200 B-scans of the test set. HDs were calculated between the predictions of convolutional neural network (CNN) 4.a and the labelings of each of the 10 human graders (top 10 rows). HDs were also calculated between CNN 4.a's predictions and the golden ground truth (GGT, 11th row) and between the GGT and the labelings of each of the 10 human graders (bottom 10 rows). Red indicates large HDs and green indicates small HDs. A color-blind-friendly version of this figure is shown in Supplementary Fig. S1.

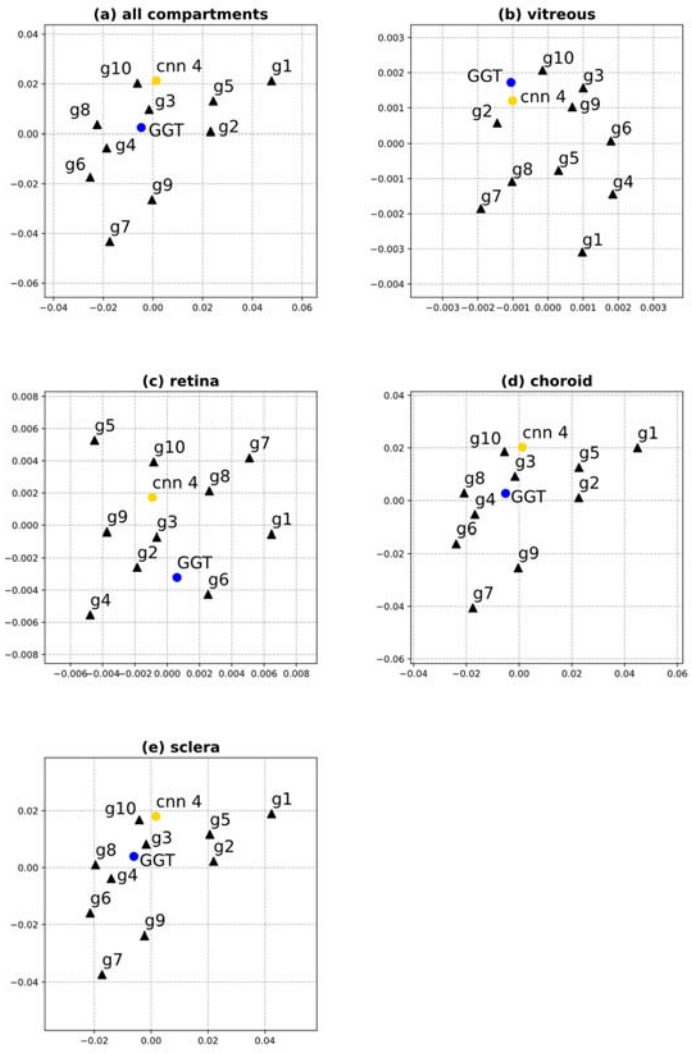


Fig. 2. Multidimensional scaling plots visualizing the mean Hamming distance (HD) among labelings of the test set by the 10 human graders and the prediction of convolutional neural network (CNN) 4.a in two-dimensional coordinate systems. The CNN was trained on ground truth annotated by all 10 human graders. Plots show mean HD across all compartments (a) and separately for each individual compartment, namely (b) vitreous, (c) retina, (d) choroid, and (e) sclera. Note the differences in the scale of the axes among the plots. Abbreviations: g = grader, GGT = golden ground truth.

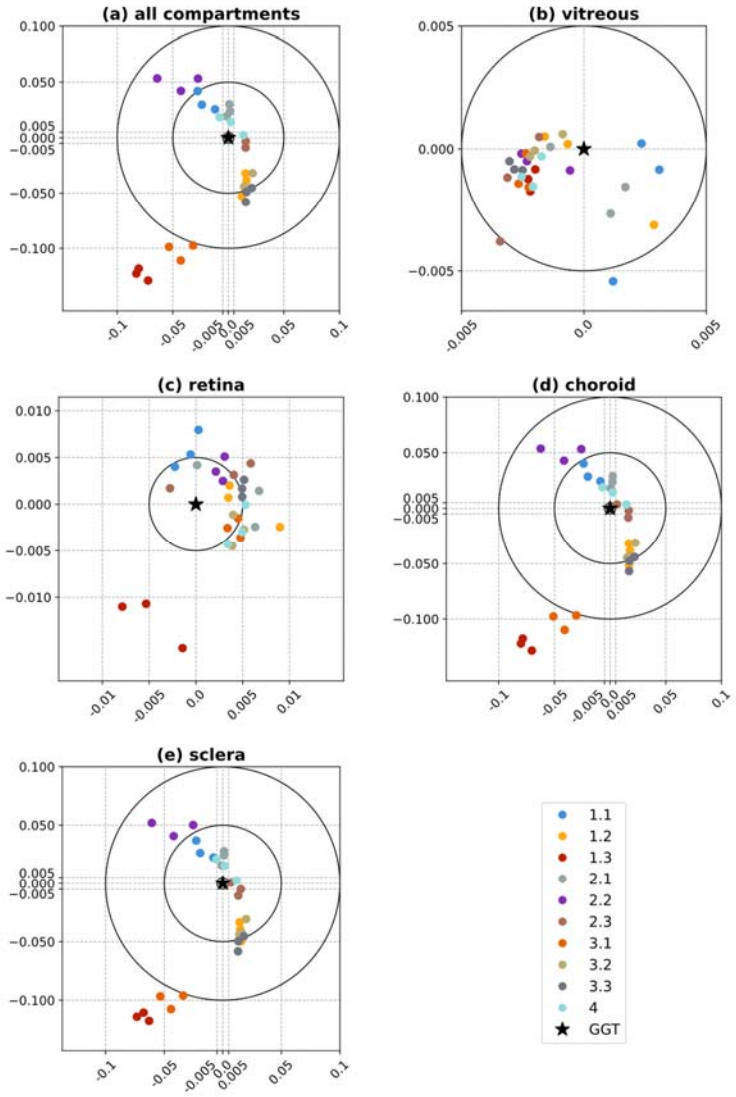


Fig. 3. Multidimensional scaling plots visualizing the mean Hamming distance (HD) among test set labels predicted by the 30 convolutional neural networks (CNNs) trained in this study. Each individual CNN (indicated as colored dot in the frame) was trained three times with the same training set. Golden ground truth is indicated as a black star and placed at the center of the coordinate systems. Plots show HD across all compartments (a) and separately for each individual compartment, namely (b) vitreous, (c) retina, (d) choroid, and (e) sclera. CNNs trained on the same base datasets are shown in the same color. Note the differences in the scale of the axes among the plots. Accessible color map from Petroff was used for data visualization.

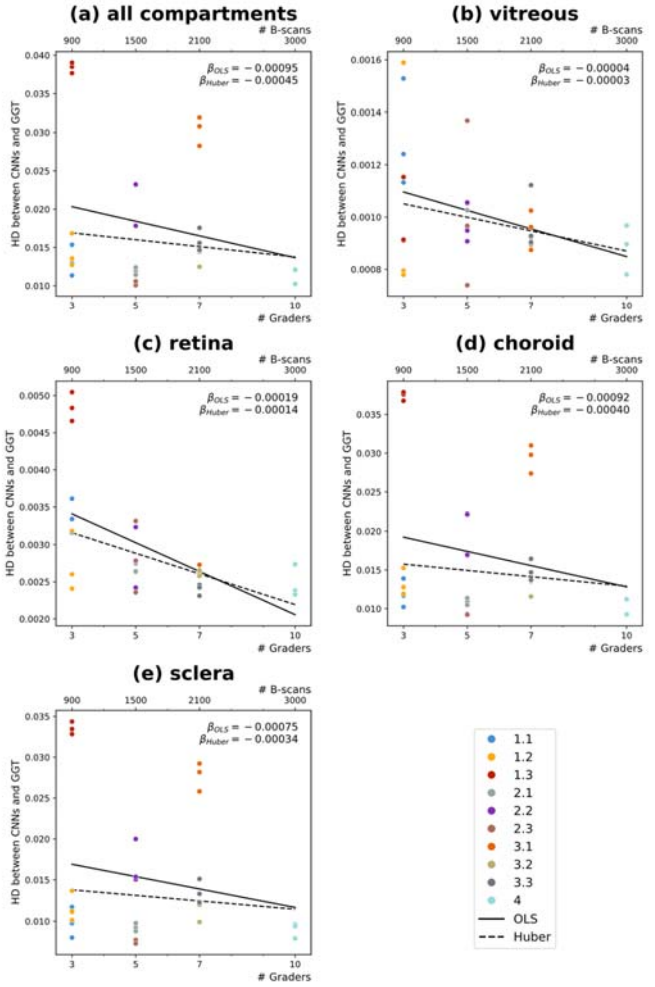


Fig. 4. Scatter plots visualizing the effect of the number of human graders included in generating a ground truth set on the predictive performance of a convolutional neural network (CNN) trained on this ground truth set. x-axis: the number of human graders that contributed to generating a ground truth set (bottom); the number of B-scans contained in a ground truth set (top). y-axis: the mean Hamming distance between a trained CNN's predictions on the test set and the golden ground truth (GGT). The GGT is used as a proxy for the "true" labels and thus measures how good a CNN's predictions are. Plots are shown across all compartments (a) and separately for individual compartments, namely (b) vitreous, (c) retina, (d) choroid, and (e) sclera. Each plot shows two linear regression lines, namely Ordinary Least Squares regression (solid line) and Huber regression (dashed line), with the regression line's slope indicated in the top-right corner as β (slopes were calculated with regard to the number of human graders). CNNs trained on the same ground truth set are shown with the same color. Please note that the scaling along the y-axis varies across the panels. Accessible color map from Petroff was used for data visualization.

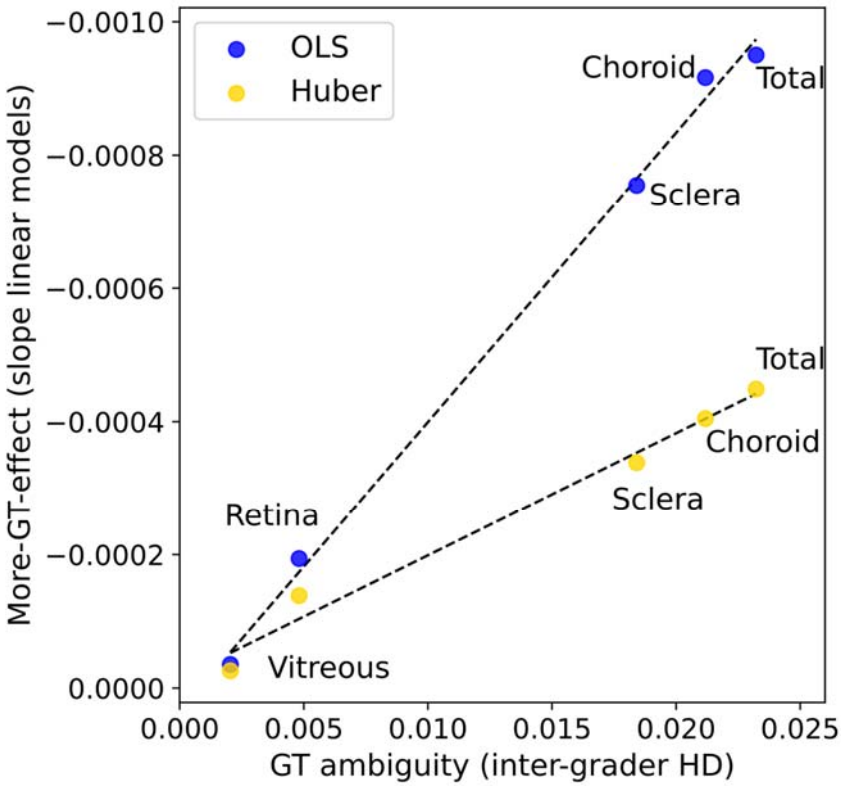


Fig. 5. The linear nature of the more-ground-truth effect." X-axis indicates ground truth ambiguity: the mean Hamming distance (HD) among labels generated by the ten human graders on the test set of 200 B-scans. . Mean HD was calculated across all four compartments (total) and for each of the four compartments separately (vitreous, retina, choroid, and sclera). Y-axis indicates slopes (β) of the regression lines from Fig. 4. Slopes indicate the strength of the effect of having more ground truth for convolutional neural network training. Slopes were calculated with respect to the number of human graders. The more negative a slope is, the stronger the effect (note that y-axis is mirrored). Yellow and blue points indicate slopes from OLS and Huber regressions, respectively. Dashed black lines are linear regressions through yellow and blue points, respectively, illustrating the linear nature of the effect.

Acknowledgements

We thank Roche, Pharma Research and Early Development (pRED), Pharmaceutical

Sciences (PS), 4070 Basel, Switzerland for financial support. The authors thank

Fabian Lutz, Flumedia Ltd., Luzern, Switzerland, for the technical support and coding of the user interface.

Author contributions statement

P.M.M.: Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing, original draft preparation, writing – review and editing, visualization, project administration.

M.P.: Conceptualization, methodology, data curation, writing, original draft preparation, writing – review and editing, visualization.

L.J.K.: Conceptualization, methodology, data curation, writing, original draft preparation, writing – review and editing, visualization.

M.R: Conceptualization, methodology, data curation, writing, original draft preparation, writing – review and editing, visualization.

L.G.: Conceptualization, methodology, data curation, writing, original draft preparation, writing – review and editing, visualization.

F.G.H.: Writing, writing – review and editing, visualization.

P.L.M.: Conceptualization, methodology, data curation, writing, original draft preparation, writing – review and editing, visualization.

P.V.: Methodology, data curation, writing, original draft preparation, writing – review and editing, visualization.

K.F.: Conceptualization, methodology, data curation, writing, original draft preparation, writing – review and editing, visualization.

P.A.K.: Writing, review and editing, visualization.

J.Z.V.: Writing, review and editing, visualization.

S.Z.: Writing, review and editing, visualization.

J.W.: Conceptualization, methodology, software, data curation, writing, original draft preparation, writing – review and editing, visualization.

P.K.: Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing, original draft preparation, writing – review and editing, visualization, project administration.

T.J.E.: Methodology, writing – review and editing, visualization.

S.P.R.: Conceptualization, methodology, data curation, writing, original draft preparation, writing – review and editing, visualization.

P.W.H.: Conceptualization, methodology, data curation, writing, original draft preparation, writing – review and editing, visualization.

M.J.: Conceptualization, methodology, data curation, writing, original draft preparation, writing – review and editing, visualization.

C.F.: Writing, review and editing, visualization.

C.E.: Writing, review and editing, visualization.

A.T.: Writing, review and editing, visualization.

H.P.N.S.: Conceptualization, writing – review and editing, visualization.

N.D.: Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing, original draft preparation, writing – review and editing, visualization, project administration.

Additional information

Competing interests

Research support was granted from Roche (Basel, Switzerland), especially with regard to data collection and the decision to publish. Roche had no role and did not interfere in the conceptualization or conduct of this study.

Authors J.W. and P.K. are salaried employees of Supercomputing Systems (Zurich, Switzerland). Authors M.J., C.F., and N.D. are salaried employees of Roche (Basel, Switzerland). P.M.M. and P.W.H. are consultants at Roche (Basel, Switzerland). Outside of the present study, the authors declare the following competing interests: P.M.M. is a consultant at Zeiss Forum and holds intellectual property for machine learning at MIMO AG and VisionAI, Switzerland. M.P. is a consultant for Apellis Pharmaceuticals (Waltham, US). F.G.H. reports research grants and consulting fees from Acucela, Allergan, Apellis, Bayer, Bioeq/Formycon, Roche/Genentech, Geuder, Heidelberg Engineering, ivericBio, Pixium Vision, Novartis, Zeiss; consulting fees from Alexion, Alzheon, Annexon, Astellas, Boehringer-Ingelheim, Cirrus, Grayburg Vision, LinBioscience, Stealth BioTherapeutics, Aerie, Oxurion. P.A.K. has been a consultant to DeepMind, Roche, Novartis, Apellis and BitFount, is a shareholder in Big Picture Medical, has received speaking fees from Heidelberg Engineering, Topcon, Allergan and Bayer, and is supported by a Moorfields Eye Charity Career Development Award (R190028A) and a UK Research & Innovation Future Leaders Fellowship (MR/T019050/1). P.L.M. received funding by the German Research Foundation (grant # MU4279/2-1). J.Z.V. has been a consultant from Alcon, Alimera Science, Allergan, Bausch & Lomb, Bayer, Brill Pharma, DORC, Novartis, and Roche; is a grant holder in Allergan and Novartis; and has been a lecturer for Topcon and Zeiss. P.V. received funding from the Swiss National Science Foundation (Grant 323530_199395) and the Janggen-Pöhn Foundation. S.Z. is a consultant for Alcon, Allergan, Bayer Healthcare Pharmaceuticals, Carl Zeiss Meditec, Endogena, Novartis Pharma AG and Roche Diagnostics and receives grant support from Bayer Healthcare Pharmaceuticals and Novartis Pharma AG. H.P.N.S. is supported by the Swiss National Science Foundation (Project funding: “Developing novel outcomes for clinical trials in Stargardt disease using structure/function relationship and deep

learning” #310030_201165 and National Center of Competence in Research Molecular Systems Engineering: “NCCR MSE: Molecular Systems Engineering (phase II)” #51NF40-182895), the Wellcome Trust (PINNACLE study), and the Foundation Fighting Blindness Clinical Research Institute (ProgStar study). H.P.N.S. is member of the scientific advisory boards of Astellas Pharma Global Development, Inc./Astellas Institute for Regenerative Medicine, Boehringer Ingelheim Pharma GmbH & Co; Gyroscope Therapeutics Ltd.; Janssen Research & Development, LLC (Johnson & Johnson); Novartis Pharma AG (CORE); Okuvision GmbH; and Third Rock Ventures, LLC. H.P.N.S. is a consultant for Gerson Lehrman Group; Guidepoint Global, LLC; and Tenpoint Therapeutics Limited. H.P.N.S. is member of the Data Monitoring and Safety Board/Committee of Belite Bio (CT2019-CTN-04690-1), ReNeuron Group Plc/Ora Inc. (NCT02464436), and F. Hoffmann-La Roche Ltd (VELODROME trial, NCT04657289; DIAGRID trial, NCT05126966) and member of the Steering Committee of Novo Nordisk (FOCUS trial; NCT03811561). All arrangements have been reviewed and approved by the University of Basel (Universitätsspital Basel, USB) and the Board of Directors of the Institute of Molecular and Clinical Ophthalmology Basel (IOB) in accordance with their conflict of interest policies. Compensation is being negotiated and administered as grants by USB, which receives them in its proper accounts. H.P.N.S. is co-director of the IOB, which is a non-profit foundation and receives funding from the University of Basel, the University Hospital Basel, Novartis, and the government of Basel-Stadt. C.E. is a consultant of Heidelberg Engineering, received together with A.T. funding from Novartis and a financial grant from the National Institute for Health Research (NIHR) Biomedical Research Centre, based at Moorfields Eye Hospital, and also from the NHS Foundation Trust and the UCL Institute of Ophthalmology. A.T. is a consultant

for Heidelberg Engineering and Optovue and has received research grant funding from Novartis and Bayer.

Conflict statement

The other authors declare no conflict.

Data availability

The source data underlying the graphs and charts presented in the main figures are available as supplementary table S1.

Code availability statement

The source code to apply the T-REX methodology described in this study is available on www.github.com/peter-maloca/T-REX and archived in Zenodo40.

References

1 Hee, M. R. *et al.* Optical coherence tomography of the human retina. *Arch. Ophthalmol.* **113**, 325-332, doi:10.1001/archopht.1995.01100030081025 (1995).

2 Fujimoto, J. & Swanson, E. The Development, Commercialization, and Impact of Optical Coherence Tomography. *Invest. Ophthalmol. Vis. Sci.* **57**, Oct1-oct13, doi:10.1167/iovs.16-19963 (2016).

3 Geevarghese, A., Wollstein, G., Ishikawa, H. & Schuman, J. S. Optical Coherence Tomography and Glaucoma. *Annu Rev Vis Sci* **7**, 693-726, doi:10.1146/annurev-vision-100419-111350 (2021).

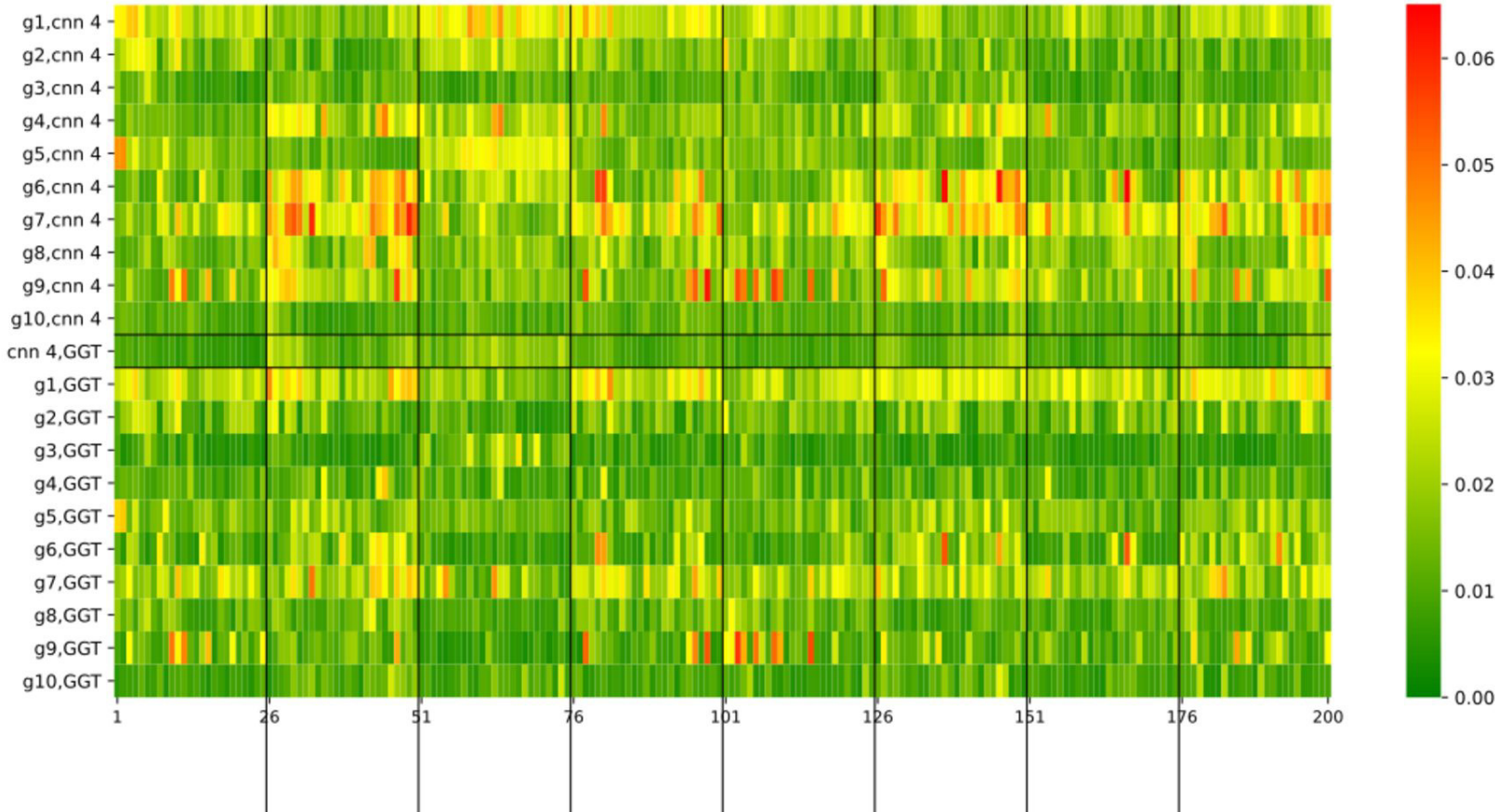
4 Litts, K. M., Zhang, Y., Freund, K. B. & Curcio, C. A. Optical coherence tomography and histology of age-related macular degeneration support mitochondria as refelctivity sources. *Retina* **38**, 445-461, doi:10.1097/iae.0000000000001946 (2018).

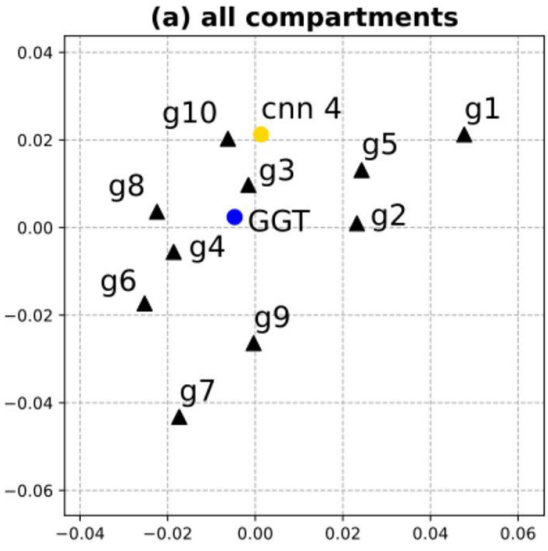
5 Waldstein, S. M. *et al.* Characterization of Drusen and Hyperreflective Foci as Biomarkers for Disease Progression in Age-Related Macular Degeneration Using Artificial Intelligence in Optical Coherence Tomography. *JAMA Ophthalmol* **138**, 740-747, doi:10.1001/jamaophthalmol.2020.1376 (2020).

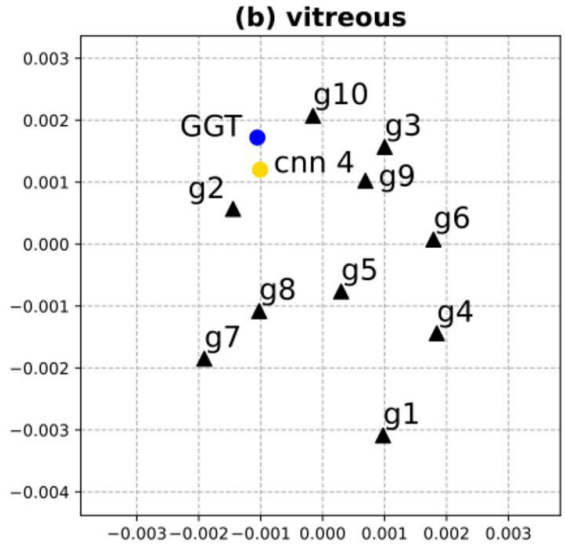
6 Sun, Z., Yang, D., Tang, Z., Ng, D. S. & Cheung, C. Y. Optical coherence tomography angiography in diabetic retinopathy: an updated review. *Eye (Lond.)* **35**, 149-161, doi:10.1038/s41433-020-01233-y (2021).

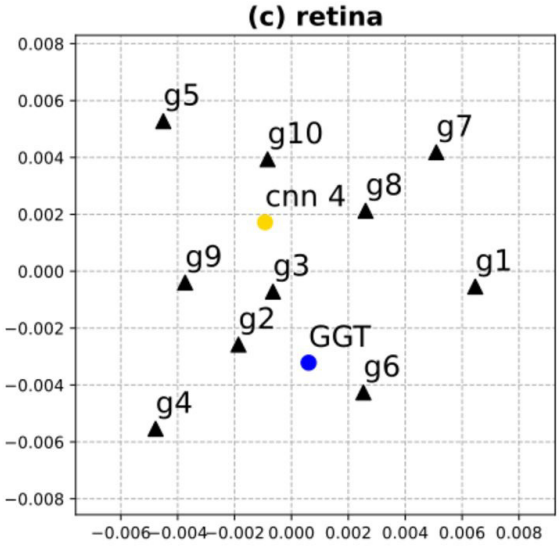
7 Johannesen, S. K., Viken, J. N., Vergmann, A. S. & Grauslund, J. Optical coherence tomography angiography and microvascular changes in diabetic retinopathy: a systematic review. *Acta Ophthalmol* **97**, 7-14, doi:10.1111/aos.13859 (2019).
8 Fujimoto, J. & Huang, D. Foreword: 25 Years of Optical Coherence Tomography. *Invest. Ophthalmol. Vis. Sci.* **57**, OCTi-OCTii, doi:10.1167/iovs.16-20269 (2016).
9 Wibbelsman, T. D. *et al.* Trends in Retina Specialist Imaging Utilization from 2012 to 2016 in the United States Medicare Fee-for-Service Population. *Am. J. Ophthalmol.* **208**, 12-18, doi:10.1016/j.ajo.2019.06.026 (2019).
10 Wibbelsman, T. D. *et al.* Reply to Comment on: Trends in Retina Specialist Imaging Utilization From 2012 to 2016 in the United States Medicare Fee-for-Service Population. *Am. J. Ophthalmol.* **211**, 229-230, doi:10.1016/j.ajo.2019.09.022 (2020).
11 Wilson, M. *et al.* Validation and Clinical Applicability of Whole-Volume Automated Segmentation of Optical Coherence Tomography in Retinal Disease Using Deep Learning. *JAMA Ophthalmol* **139**, 964-973, doi:10.1001/jamaophthalmol.2021.2273 (2021).
12 Zhou, L., Li, Q., Huo, G. & Zhou, Y. Image Classification Using Biomimetic Pattern Recognition with Convolutional Neural Networks Features. *Comput. Intell. Neurosci.* **2017**, 3792805, doi:10.1155/2017/3792805 (2017).
13 Liefers, B. *et al.* Quantification of Key Retinal Features in Early and Late Age-Related Macular Degeneration Using Deep Learning. *Am. J. Ophthalmol.* **226**, 1-12, doi:10.1016/j.ajo.2020.12.034 (2021).
14 Klimscha, S. *et al.* Spatial Correspondence Between Intraretinal Fluid, Subretinal Fluid, and Pigment Epithelial Detachment in Neovascular Age-Related Macular Degeneration. *Invest. Ophthalmol. Vis. Sci.* **58**, 4039-4048, doi:10.1167/iovs.16-20201 (2017).
15 Maloca, P. M. *et al.* Validation of automated artificial intelligence segmentation of optical coherence tomography images. *PLoS One* **14**, e0220063, doi:10.1371/journal.pone.0220063 (2019).
16 Ronneberger, O., Fischer, P. & Brox, T. 234-241 (Springer International Publishing).
17 Lee, C. S., Baughman, D. M. & Lee, A. Y. Deep learning is effective for the classification of OCT images of normal versus Age-related Macular Degeneration. *Ophthalmol Retina* **1**, 322-327, doi:10.1016/j.oret.2016.12.009 (2017).
18 Ting, D. S. W. *et al.* Artificial intelligence and deep learning in ophthalmology. *Br. J. Ophthalmol.* **103**, 167-175, doi:10.1136/bjophthalmol-2018-313173 (2019).
19 Tjoa, E. & Guan, C. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems* **32**, 4793-4813 (2021).
20 Holzinger, A., Biemann, C., Pattichis, C. S. & Kell, D. B. What do we need to build explainable AI systems for the medical domain? *ArXiv* **abs/1712.09923** (2017).
21 Samek, W. & Müller, K.-R. in *Explainable AI: interpreting, explaining and visualizing deep learning* 5-22 (Springer, 2019).
22 Maloca, P. M. *et al.* Unraveling the deep learning gearbox in optical coherence tomography image segmentation towards explainable artificial intelligence. *Commun Biol* **4**, 170, doi:10.1038/s42003-021-01697-y (2021).

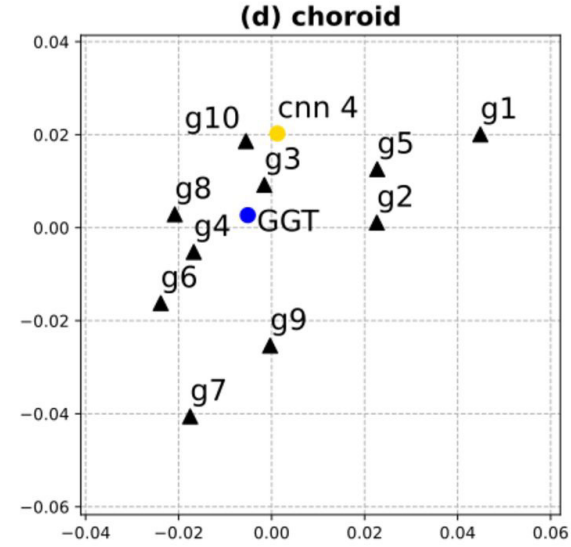
23 Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* **533**, 452-454, doi:10.1038/533452a (2016).
24 Brady, C. J., Mudie, L. I., Wang, X., Guallar, E. & Friedman, D. S. Improving Consensus Scoring of Crowdsourced Data Using the Rasch Model: Development and Refinement of a Diagnostic Instrument. *J. Med. Internet Res.* **19**, e222, doi:10.2196/jmir.7984 (2017).
25 Petroff, M. *Accessible Color Cycles for Data Visualization*. (2021).
26 De Fauw, J. *et al.* Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **24**, 1342-1350, doi:10.1038/s41591-018-0107-6 (2018).
27 Gerendas, B. S., Bogunovic, H. & Schmidt-Erfurth, U. Deep Learning-Based Automated Optical Coherence Tomography Segmentation in Clinical Routine: Getting Closer. *JAMA Ophthalmol* **139**, 973-974, doi:10.1001/jamaophthalmol.2021.2309 (2021).
28 Panch, T., Mattie, H. & Celi, L. A. The "inconvenient truth" about AI in healthcare. *NPJ Digit Med* **2**, 77, doi:10.1038/s41746-019-0155-4 (2019).
29 Grzybowski, A. & Brona, P. Analysis and Comparison of Two Artificial Intelligence Diabetic Retinopathy Screening Algorithms in a Pilot Study: IDx-DR and Retinalyze. *J Clin Med* **10**, doi:10.3390/jcm10112352 (2021).
30 Renard, F., Guedria, S., Palma, N. & Vuillerme, N. Variability and reproducibility in deep learning for medical image segmentation. *Sci. Rep.* **10**, 13724, doi:10.1038/s41598-020-69920-0 (2020).
31 Gunning, D., Explainable artificial intelligence (xAI). *Technical Report, Defense Advanced Research Projects Agency (DARPA)* (2017).
32 Holzinger, A., Kieseberg, P., Weippl, E. & Tjoa, A. M. . Current Advances, Trends and Challenges of Machine Learning and Knowledge Extraction: From Machine Learning to Explainable AI. . *Springer Lecture Notes in Computer Science LNCS 11015*, 1-8, doi:10.1007/978-3-319-99740-7-1 (2018).
33 Matheny, M. E., Whicher, D. & Thadaney Israni, S. Artificial Intelligence in Health Care: A Report From the National Academy of Medicine. *JAMA* **323**, 509-510, doi:10.1001/jama.2019.21579 (2020).
34 Armstrong, J. S. in *Principles of Forecasting: A Handbook for Researchers and Practitioners* (ed J. Scott Armstrong) 417-439 (Springer US, 2001).
35 Raffel, C. *et al.* Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **21**, 140:141-140:167 (2020).
36 Wang, C. *et al.* in *AACL.*
37 Jia, Y. *et al.* in *INTERSPEECH.*
38 Kapoor, S. & Narayanan, A. *Leakage and the Reproducibility Crisis in ML-based Science*. (2022).
39 National Academies of Sciences, E. & Medicine. Reproducibility and replicability in science. (2019).
40 Bishop, C. M. & Nasrabadi, N. M. Pattern Recognition and Machine Learning. *J. Electronic Imaging* **16**, 049901 (2007).
41 Hastie, T. J., Tibshirani, R. & Friedman, J. H. in *Springer Series in Statistics.*
42 Ronneberger, O. F., P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *CoRR* (2015).
43 Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. (Springer New York, 2009).

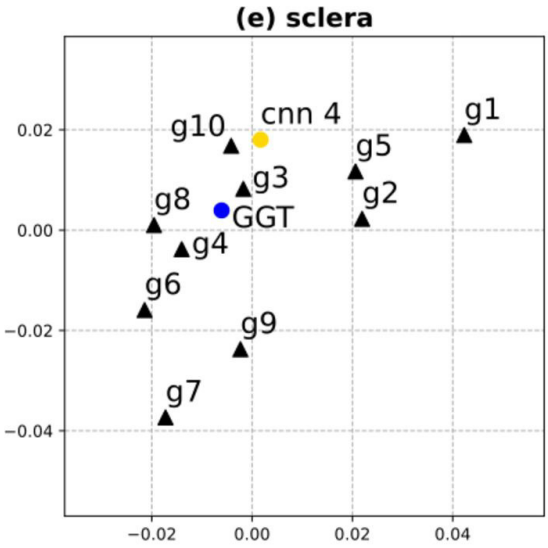

g1,cnn 4
g2,cnn 4
g3,cnn 4
g4,cnn 4
g5,cnn 4
g6,cnn 4
g7,cnn 4
g8,cnn 4
g9,cnn 4
g10,cnn 4
cnn 4,GGT
g1,GGT
g2,GGT
g3,GGT
g4,GGT
g5,GGT
g6,GGT
g7,GGT
g8,GGT
g9,GGT
g10,GGT
1
26
51
76
101
126
151
176
200
0.06
0.05
0.04
0.03
0.02
0.01
0.00

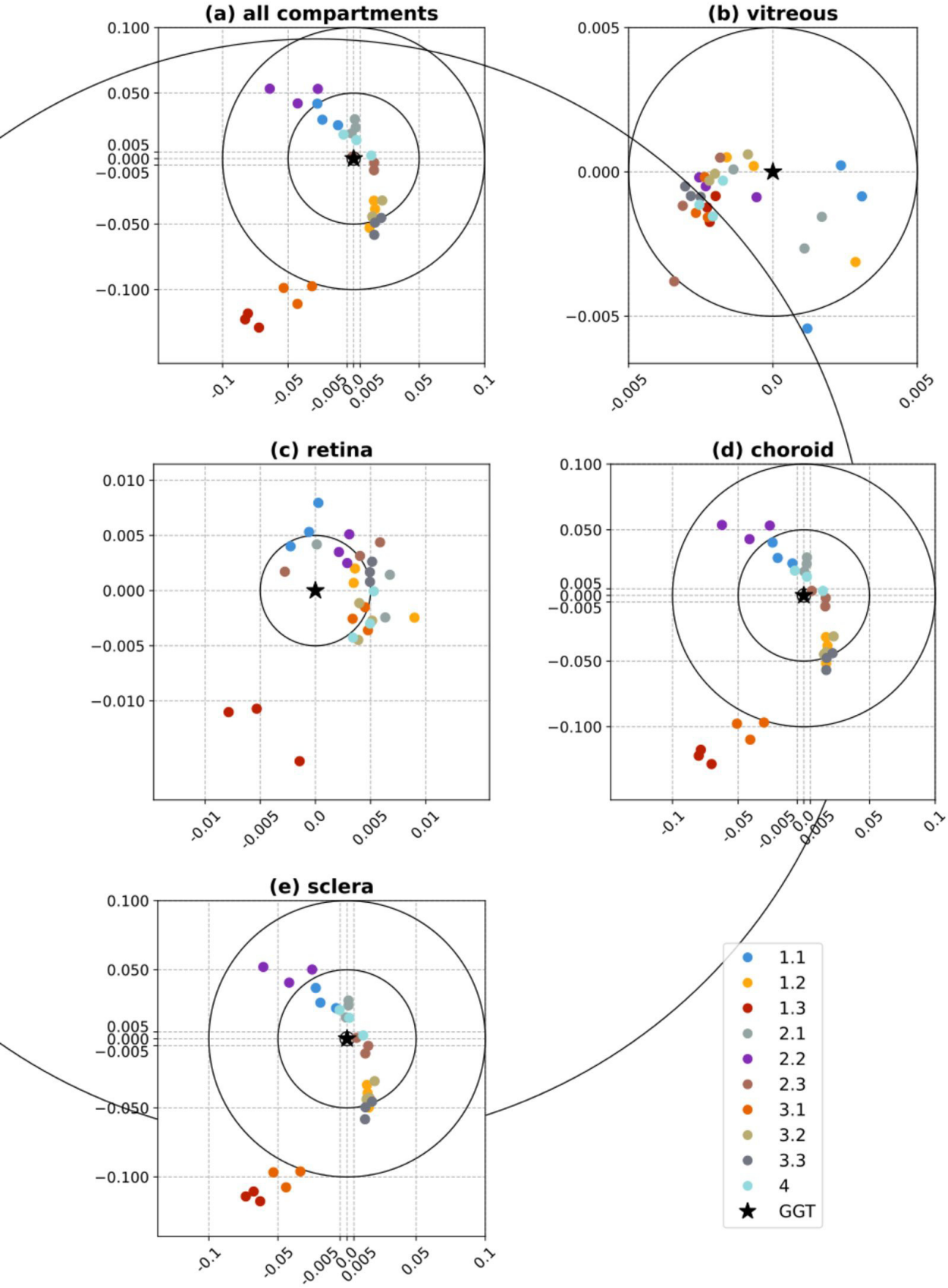

(a) all compartments
g10
cnn 4
g1
g5
g8
g3
GGT
g2
g4
g6
g9
g7

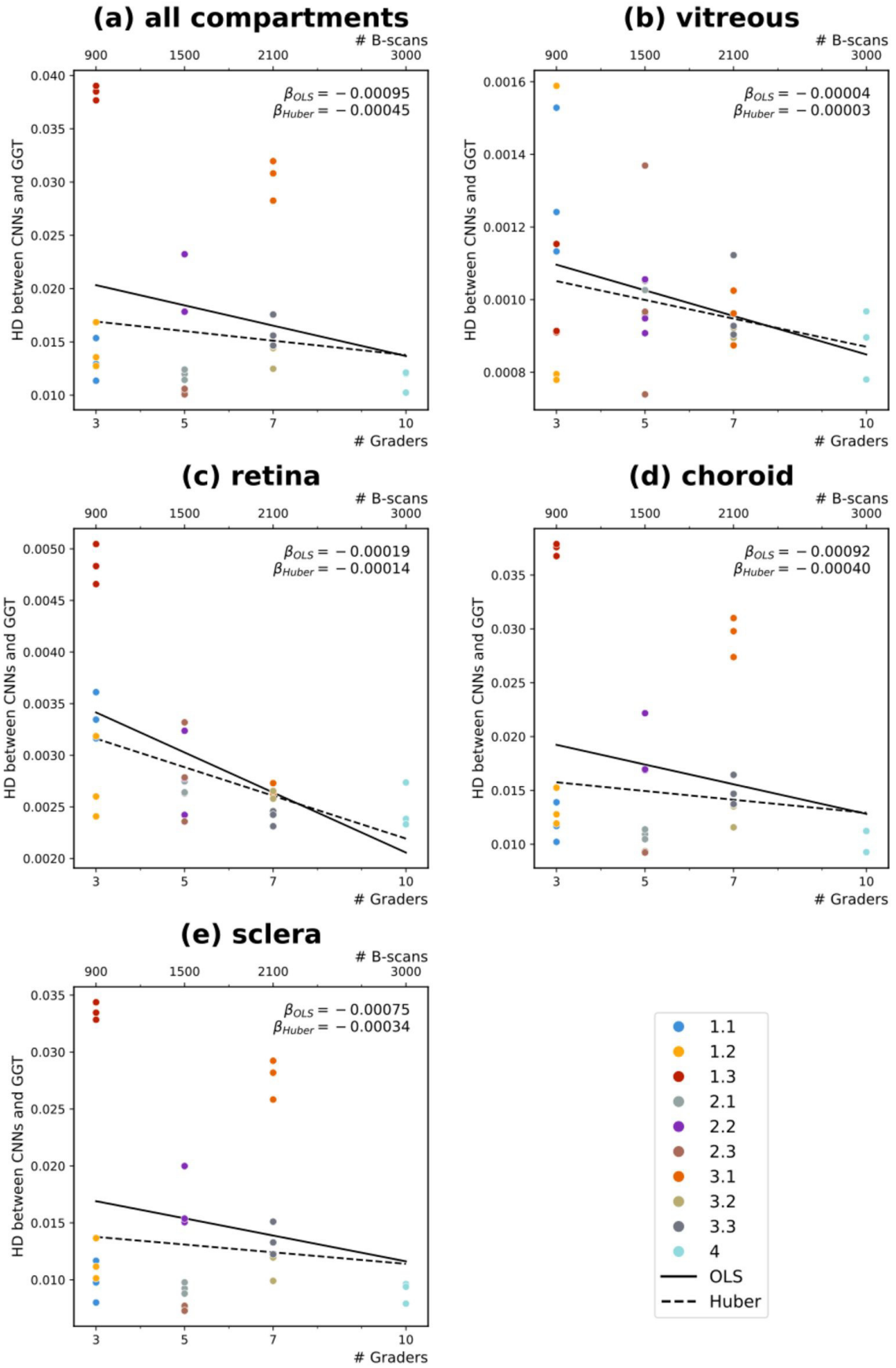

(b) vitreous
g10
GGT
g3
cnn 4
g9
g2
g6
g5
g8
g4
g7
g1

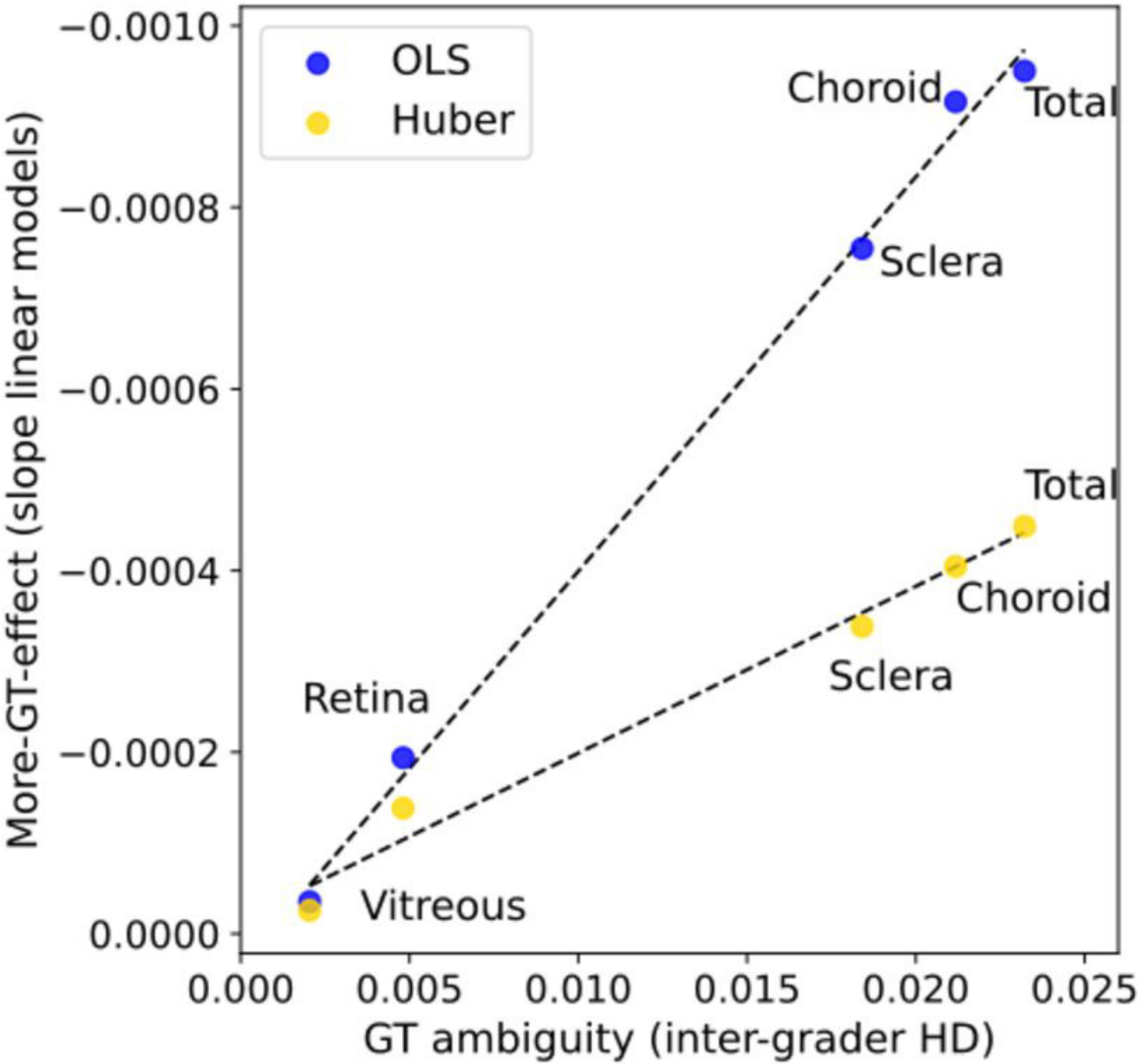

(c) retina
g5
g10
g7
cnn 4
g8
g9
g3
g1
g2
GGT
g6
g4


(d) choroid
g10
cnn 4
g1
g3
g5
g8
g2
g4
GGT
g6
g9
g7


(e) sclera
g10
cnn 4
g1
g5
g3
g8
GGT
g2
g4
g6
g9
g7


(a) all compartments
(b) vitreous
(c) retina
(d) choroid
(e) sclera
1.1
1.2
1.3
2.1
2.2
2.3
3.1
3.2
3.3
4
GGT


(a) all compartments
B-scans
900
1500
2100
3000
$\beta_{OLS} = -0.00095$
$\beta_{Huber} = -0.00045$
HD between CNNs and GGT
0.040
0.035
0.030
0.025
0.020
0.015
0.010
3
5
7
10
Graders
(b) vitreous
B-scans
900
1500
2100
3000
$\beta_{OLS} = -0.00004$
$\beta_{Huber} = -0.00003$
HD between CNNs and GGT
0.0016
0.0014
0.0012
0.0010
0.0008
3
5
7
10
Graders
(c) retina
B-scans
900
1500
2100
3000
$\beta_{OLS} = -0.00019$
$\beta_{Huber} = -0.00014$
HD between CNNs and GGT
0.0050
0.0045
0.0040
0.0035
0.0030
0.0025
0.0020
3
5
7
10
Graders
(d) choroid
B-scans
900
1500
2100
3000
$\beta_{OLS} = -0.00092$
$\beta_{Huber} = -0.00040$
HD between CNNs and GGT
0.035
0.030
0.025
0.020
0.015
0.010
3
5
7
10
Graders
(e) sclera
B-scans
900
1500
2100
3000
$\beta_{OLS} = -0.00075$
$\beta_{Huber} = -0.00034$
HD between CNNs and GGT
0.035
0.030
0.025
0.020
0.015
0.010
3
5
7
10
Graders
1.1
1.2
1.3
2.1
2.2
2.3
3.1
3.2
3.3
4
OLS
Huber


OLS
Huber
Choroid
Total
Sclera
Total
Choroid
Sclera
Retina
Vitreous
−0.0010
−0.0008
−0.0006
−0.0004
−0.0002
0.0000
0.000
0.005
0.010
0.015
0.020
0.025
More-GT-effect (slope linear models)
GT ambiguity (inter-grader HD)