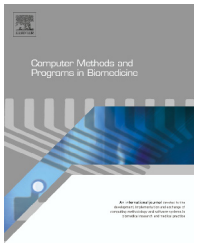
COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 130 (2016) 93–105

journal homepage: www.intl.elsevierhealth.com/journals/cmpb



A novel benchmark model for intelligent annotation of spectral-domain optical coherence tomography scans using the example of cyst annotation☆

Ehsan Shahrian Varnousfaderani, Jing Wu, Wolf-Dieter Vogl, Ana-Maria Philip, Alessio Montuoro, Roland Leitner, Christian Simader, Sebastian M. Waldstein, Bianca S. Gerendas, Ursula Schmidt-Erfurth*

Christian Doppler Laboratory for Ophthalmic Image Analysis, Vienna Reading Center, Department of Ophthalmology, Medical University of Vienna, Vienna, Austria

ARTICLE INFO

Article history:
Received 21 August 2015
Received in revised form
5 February 2016
Accepted 10 March 2016

Keywords:
Benchmark dataset
Cyst segmentation
SD-OCT

ABSTRACT

Background and objectives: The lack of benchmark data in computational ophthalmology contributes to the challenging task of applying disease assessment and evaluate performance of machine learning based methods on retinal spectral domain optical coherence tomography (SD-OCT) scans. Presented here is a general framework for constructing a benchmark dataset for retinal image processing tasks such as cyst, vessel, and subretinal fluid segmentation and as a result, a benchmark dataset for cyst segmentation has been developed.

Method: First, a dataset captured by different SD-OCT vendors with different numbers of scans and pathology qualities are selected. Then a robust and intelligent method is used to evaluate performance of readers, partitioning the dataset into subsets. Subsets are then assigned to complementary readers for annotation with respect to a novel confidence based annotation protocol. Finally, reader annotations are combined based on their performance to generate final annotations.

Result: The generated benchmark dataset for cyst segmentation comprises 26 SD-OCT scans with differing cyst qualities, collected from 4 different SD-OCT vendors to cover a wide variety of data. The dataset is partitioned into three subsets which are annotated by complementary readers based on a confidence based annotation protocol. Experimental results show annotations of complementary readers are combined efficiently with respect to their performance, generating accurate annotations.

Conclusion: Our results facilitate the process of generating benchmark datasets. Moreover the generated benchmark data set for cyst segmentation can be used reliably to train and test machine learning based methods.

☆ *Grant*: Austrian Federal Ministry of Science, Research and Economy; National Foundation for Research, Technology and Development.
* *Corresponding author*. Tel.: +43 14040079310.
E-mail address: bianca.gerendas@meduniwien.ac.at (B.S. Gerendas).
http://dx.doi.org/10.1016/j.cmpb.2016.03.012
0169-2607/

1. Introduction

In the areas of ophthalmic disease assessment, diagnosis, and treatment planning, analysis of retinal morphology plays an important role [1–5]. The analysis of retinal cysts, subretinal fluid, and fluid under the retinal pigment epithelium holds the key to understanding and treating diseases such as age-related macular degeneration (AMD) [6–8], diabetic macular edema (DME) [9] (Gerendas, unpublished data, 2014) and macular edema due to retinal vein occlusion (MEVO) (Waldstein, unpublished data, 2014). Patients are imaged using spectral-domain optical coherence tomography (SD-OCT), a non-invasive modality for acquiring high resolution, 3D cross sectional volumetric images of the retina and the sub-retinal layers [10,11].

Today, SD-OCT is the most important ancillary test for the diagnosis of sight threatening diseases [12]. For these reasons, understanding and measuring the position, size, and composition of cysts is needed, for example. Whereas the delineation of subretinal fluid can be performed easier, the ability for manual delineation of cysts is limited due to the difficulty for human experts to accurately and reproducibly identify them. In addition, when the number of cysts in an image is large, manual identification becomes tedious or even impossible, thus automated systems are preferable.

However, such systems require large volumes of annotated and variable data for training and testing. Until now, there is no publically available dataset featuring expertly annotated and investigated multi-vendor retinal cysts usable as ground truth for the development of automated or semi-automated cyst segmentation methods. As a result, it is difficult to compare methods that have been developed in this area. Such datasets have been created for other purposes in the form of the DRIVE database of digital retinal images for vessel extraction [13], the REVIEW database for the measurement of retinal vessel widths [14], and the STARE database [15] for vessel segmentation, and the database for the coronary artery algorithm evaluation framework [16].

The primary focus of this paper is the development of a novel intelligent method of reader (defined as all people who perform manual annotations for cyst delineation) evaluation, task assignment, and data partitioning of retinal SD-OCT scans based on cyst annotation and the construction of a benchmark dataset for use in training and testing. The data set used here features retinal cysts obtained from 4 of the major SD-OCT scanner vendors, annotated by expert readers at the Vienna Reading Center (VRC) and a subgroup, specially trained for cyst annotation from the Christian Doppler Laboratory for Ophthalmic Image Analysis (OPTIMA) and the evaluation of said annotations. The cysts in question have been manually annotated using a proprietary annotation tool developed for this purpose at OPTIMA. In addition, the reader evaluation and task assignment system has been designed to be applicable for other retinal pathologies other than cysts.

Previous methods of (semi-)automated cyst segmentation [9,17] have primarily used local or private datasets featuring a limited number of scans and/or only a single vendor. In addition, annotations of the cysts have been often carried out by a single non-expert reader without validated reading tools which may result in subjectivity bias or reproducibility errors. Our goal to develop a benchmark dataset of cyst annotations in SD-OCT will allow further development in (semi-)automated cyst segmentation algorithms with the aid of a fully annotated and validated dataset as well as the ability to compare developed methods using a single reference standard. For this purpose, a reader evaluation and task assignment system have been developed to evaluate the annotations of 3 readers and 1 expert supervisor on a dataset consisting of 26 SD-OCT scans from 4 major vendors taken from the VRC and OPTIMA database of patients suffering from AMD and glaucoma. Evaluation of the resulting annotations is performed, along with an intense reader evaluation to judge reader performance.

This paper is organized in 4 sections. In Section 2, the primary method for task assignment, annotation, data evaluation, and reader evaluation is described in detail. In Section 3, results of the annotation process are presented, and summarized and discussed in Section 4.

2. Methods

In this paper, we propose a general framework for constructing benchmark datasets for retinal image processing tasks such as cyst, vessel, and subretinal fluid segmentation as shown in Fig. 1 and as a result, a benchmark dataset for cyst segmentation is developed. The main contributions of this paper are the novel *data selection*, *task assignment*, and *annotation combination*

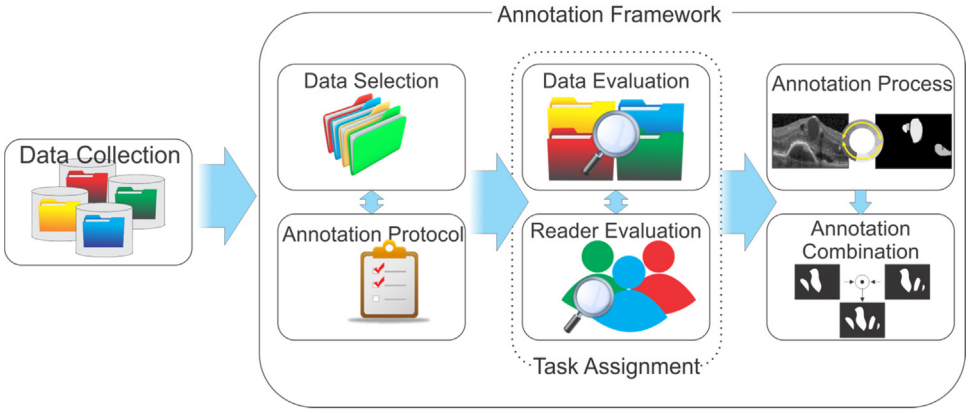


Fig. 1 – Illustration of proposed annotation framework.

Table 1 – Selected SD-OCT volume scans from different vendors.

SD-OCT scanner vendor	Number of SD-OCT scans	Size (pixel)	Depth resolution (μm/pixel)	Spatial resolution (μm/pixel)
Zeiss Cirrus	9	1024 × 512 × 200	2	12
Nidek RS-3000	11	512 × 475 × 128	4	13
Heidelberg Spectralis	3	496 × 512 × 49	4	12
Topcon 3D-OCT-2000	3	885 × 769 × 256	2	8

stages which are explained in detail in the following section. All readers and supervisors within this study are from a well-controlled and standardized setting of the Vienna Reading Center being one of the world leading institutions for retinal image evaluation.

2.1. *Dataset selection*

The VRC and OPTIMA have collected many datasets from different diseases and SD-OCT vendors in different image quality throughout many clinical studies. For this work, 26 macular SD-OCT scans, each 6 mm in length and 2 mm in depth, have been selected from different vendors comprising 9, 3, 3, and 11 scans from Zeiss Cirrus, Nidek RS-3000, Heidelberg Spectralis, and Topcon 3D-OCT-2000, respectively. The SD-OCT of the Zeiss Cirrus instrument, for example, has 1024 pixels (2 mm) in depth and 512 pixels in length (6 mm) and thus has a depth resolution of 2 μm/pixel and spatial resolution of 12 μm/pixel. The size, depth and spatial resolutions of all other SD-OCT scans are shown in Table 1.

Sample slices of data from different vendors are shown in Fig. 2. This dataset not only includes SD-OCT volumes from multiple vendors but also features different levels of cyst pathology. The dataset has been selected by an expert consortium of ophthalmologists evaluating pathology together with an expert consortium of medical image specialists to judge image quality to reflect a real world dataset.

Due to the differences in the scanners mentioned previously, retinal scans acquired vary in appearance. As a result, the appearance of the same pathology can also vary. For the purposes of performing reader evaluation and task assignment, the system must be trained using scans from all vendors as a given dataset may feature any combination of vendor scans.

The 26 SD-OCT scans for development of this benchmark system have been chosen to include a variety of image qualities, ranging from noisy to clean. In addition, the level of pathology varies across a similar scale ranging from low to high pathology presence. In addition, the quality of the cysts is also important, ranging from structures with high cyst likelihood to low cyst likelihood, while still identified as cysts.

2.2. *Annotation protocol*

The cyst annotation protocol reduces the level of inter- and intra-reader variability when readers annotate cyst regions on SD-OCT images. As a result, the annotations obtained can be used as a benchmark for training and validation of cyst segmentation methods. The key stage of cyst annotation is correct identification of the cyst regardless of the location within the image or the position within any particular retinal layer. Thus readers are trained to use the following requirements for cyst identification:

- *General*: Cyst regions shall only be marked when the reader is confident that the region represents a cyst. Regions with possible cyst representation with low probability and without confidence shall not be marked within this annotation protocol.
- *Distinction*: Cyst regions tend to have visible boundaries and clear distinction from non-cyst regions.
- *Shape*: Cysts tend to have a circular/oval shape, however combinations of nearby cysts may lead to a cystic region with other shapes.
- *Continuity*: Cysts are usually present in consecutive B-scans, therefore readers should check for cyst presence in the following and previous B-scans; the Continuity criteria are valid for high density scans with density resolution less than 60 μm/pixel and readers should be aware that some vendors have wider spaced B-scans (low density scans with density resolution greater than 60 μm/pixel) that may affect the continuity (e.g. Heidelberg Spectralis).
- *Data pool*: The Reader must use all available imaging planes (*XY*, *XZ* and *YZ*) to determine if the cyst is true. A cyst

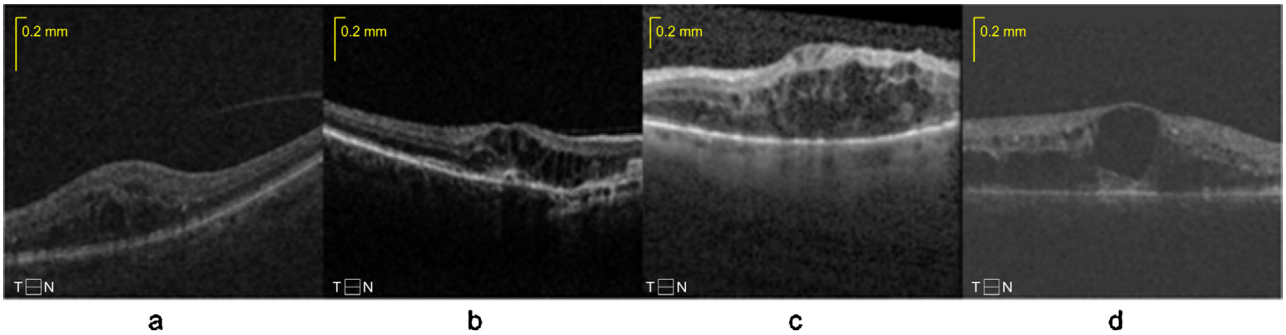


Fig. 2 – Examples of images from each of the 4 scanner vendors from (a) Cirrus, (b) Nidek, (c) Spectralis and (d) Topcon.

must be visible not only in the plane used in regular clinical practice (*XZ*) but also in the other planes (*XY* or *YZ*)
- *Position*: From a clinical perspective, it is highly unlikely to see a cyst in the retinal layers below the external limiting membrane (ELM), therefore the reader should evaluate potential cysts in these regions with more caution; cysts close to the fovea are more likely to adversely affect vision, therefore readers are asked to pay more attention to identify such cysts.

In the event that a cyst with a speculative boundary is encountered, the reader is asked to annotate the boundary as carefully and accurately as possible at the inner speculative border using best judgment to infer the boundary. Readers assign a cyst confidence to annotated regions which numerically quantifies the confidence of a selected region being a cyst based on the reader's opinion. The cyst confidence has a range [1–5] in which 1 and 5 correspond to low and high confidences, respectively.

2.3. *Task assignment procedure*

Task assignment describes a process of assigning clinical data to a reader for the annotation of a particular pathology. Clinical studies usually deal with large datasets but limited number of readers. Data annotation by a single reader is time consuming and biased toward that particular reader. This bias can be reduced by annotating data with different readers but it increases the time requirement drastically. For example, if the average time for cyst annotation on a single B-scan is 15 min, then for a single 200 B-scan volume it is 50 h. Thus for a simple study comprised of 20 SD-OCT volumes, this time increases to 1000 h. If a reader works for 40 h a week, it will take more than 6 months to collect annotations – and this does not consider that a reader will never be able to annotate cysts for 40 h a week due to concentration problems when annotating for 8 h consecutively.

A solution to reduce annotation time is to partition data into smaller subsets and to annotate them with different readers and then combine the results to generate the final annotation. This solution is simple but there are a few concerns which should be taken into account. Firstly, how can the data be split into subsets without causing selection bias? Secondly, which reader should annotate which subset to prevent reader annotation bias?

The proposed task assignment system removes selection biases toward vendors and pathology at the data evaluation stage by splitting the dataset into subsets comprised of data from different vendors and different pathology quality. Moreover, reader annotation bias is removed in the reader evaluation stage by annotating a given subset of data with complementary readers, which are combined with respect to their performance. The overall data and reader evaluation process, as well as the task assignment procedure can be seen in Fig. 3, where each color represents a similar group of either data or readers and each distinct hue represents an element within each range.

2.3.1. *Data evaluation*

The dataset is divided into small partitions at the data evaluation stage by applying two level hierarchical clustering as shown in Fig. 3a. In this example, the SD-OCT volumes of four different vendors are shown (Fig. 3a) in different colors (light green, purple, blue, and intense green). Furthermore, three

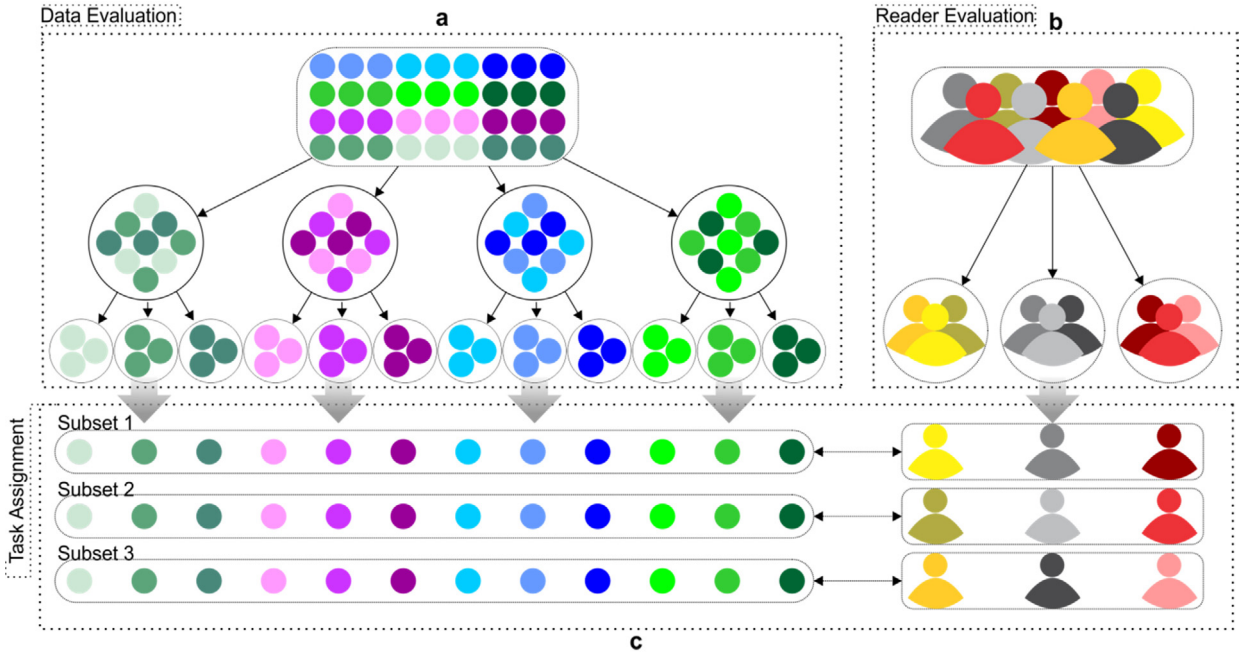


Fig. 3 – Illustration of task assignment procedure, (a) data evaluation partitions the selected data into small partitions, (b) reader evaluation finds the complementary readers, and (c) task assignment assigns subsets of data to complementary readers for annotation.

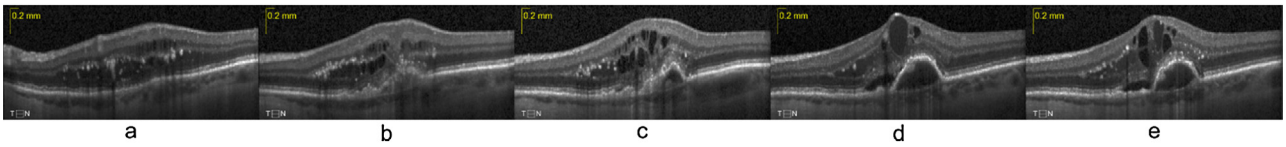


Fig. 4 – 5 Exemplar B-scans from a Heidelberg Spectralis SD-OCT scan.

different color hues represent the pathology quality. Level one shows how the data is clustered into four vendor partitions. Each vendor partition is further clustered into three smaller partitions with respect to pathology quality. Clustering is performed in such a way to minimize/maximize the variability within/between small partitions. The generated small partitions in the second level indicate data with different pathology qualities for specific vendors. Finally, samples are randomly selected from each small partition at level 2 independently and combined to build a given data subset. This results in at least one sample from each of the partitions present in a data subset even when the probability of being selected is small. Three generated subsets are shown in Fig. 3c. The number of selected samples from each small partition at level 2 depends on the size of the subsets and the proportional size of the small partition to that of the dataset, thus mimicking the original dataset on a smaller scale.

The pathology quality of cysts is calculated with the aid of the readers in this experiment. Three experienced readers are asked to rank cyst quality in each SD-OCT scan as high, medium, and low. High pathology quality relate to cysts that are clearly visible and have high distinction from non-cystic objects, implying ease of reader annotation. Low pathology quality of cyst means that the boundaries of the cysts are not clearly visible and it will be difficult for the reader to identify and annotate these, usually due to low distinction and high noise. The overall pathology quality of cyst for an SD-OCT volume is computed as the median of the readers' ranks.

Thus our dataset of 26 SD-OCT scans is divided into the 4 vendor partitions at the first level, followed by division into smaller partitions with respect to three different degrees of cyst quality, high, medium, and low. As a result, 12 small partitions are generated at the second level. Three subsets of data are generated by combining randomly selected (without replacement) samples from each small partition, comprising volumes from different vendors with different pathology qualities and therefore remove annotation biases toward vendors and cyst quality.

2.3.2. *Reader evaluation*

Annotation of pathologies by a single reader is time consuming and biases the results toward that particular reader, especially troublesome if the reader fails to identify low confidence cysts, then information is lost. This problem can be resolved by complementary reader annotation. Complementary readers may not be the most accurate but they must not suffer from the same negative annotation characteristics, such as poor annotation of certain types or qualities of cysts. Thus it is necessary to determine which readers are complementary to one another as well as how to combine the annotations of complementary readers.

To this end, we evaluate reader performance using a cyst evaluation set: Five B-scans (the slices comprising an OCT volume) from every SD-OCT volume in our dataset are expertly selected ensuring that the B-scans contain cysts of low to high quality. This is exemplified in Fig. 4, where 5 B-scans from Heidelberg Spectralis show low cyst quality in Fig. 4a, improving in cyst quality from left to right culminating in Fig. 4e. This cyst evaluation set is annotated twice by each reader (and one expert supervisor) following the annotation protocol defined in Section 2.2 and with a 1 day cool off between the first and second annotations.

2.3.2.1. *Ordinal confidence to interval confidences.* The assigned cyst confidences are of ordinal data type and as a result may not feature even intervals, thus making it impossible to apply any arithmetic operation on said ordinal data [18,19] or to compute any distance measure. Thus an ordinal to interval conversion is required, the simplest being conversion of ordinal data into two groups, yes or no, in relation to the cost of losing ordinal information [20]. Equal Distance Scoring (EDS) makes an equal interval assumption [21], although it is not efficient when ranks are not uniformly distributed [22]. Monotonic random scoring (MRS) uses random numbers in place of the ordinal scale but may lead to different intervals from underlying ordinal intervals [18]. Overall, several complex conversion methods have been proposed but the perfect ordinal-to-interval conversion is still controversial [23,24].

The underlying intervals of ordinal cyst confidences for each reader as well as for the expert supervisor are found by analyzing correlations between cyst confidences and cyst distinctions. Cyst distinction indicates how different a cyst is from its surrounding background and it is inversely proportional to the overlap of intensity histograms of the region inside and the region in the outer boundary of the cyst. The distinction of the cyst i is defined as:

$$D_i = 1 - \sum_{j=1}^{d} \min(H_j^{inner}, H_j^{outer\ boundary}) \quad (1)$$

where H^{inner} and $H^{outer\ boundary}$ are normalized intensity histograms of the region inside and the region in the outer boundary of cyst i. The index j represents histogram bins and d is the number of bins in the histogram. In our experiment, outer boundary region is considered as a 15 pixels wide region around the object that is computed using morphological dilation and subtraction and 25 bins histogram ($d = 25$) is used to compute the cyst distinction. The example of inner and outer boundary regions is shown in Fig. 5b with red and blue colors, respectively.

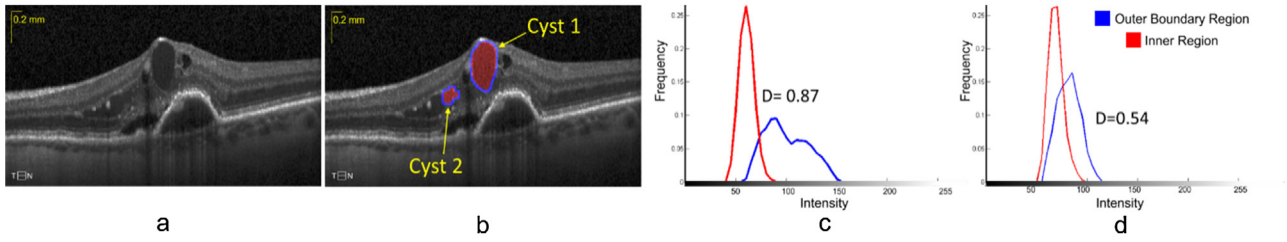


Fig. 5 – Distinction score illustration. (a) B-scan from SD-OCT scan from Heidelberg Spectralis. (b) Two cysts with highlighted inner and outer boundary regions. (c, d) Histograms of inner regions (red) and outer regions (blue) of cyst 1 and cyst 2. (For interpretation of the references to color in figure legend, the reader is referred to the web version of the article.)

The distribution of cyst distinction scores for different cyst confidences can be computed for readers and expert supervisor using their annotations on the cyst evaluation set. The subjectivity of readers and supervisor result in different cyst confidences being assigned particular distinction, this leads to different distributions of cyst distinction scores by readers and supervisor for different cyst confidences.

In order for uniform subjectivity, distinction distributions of readers are linearly transformed with respect to the reference distinction distributions of the supervisor. For given data X with the mean μ and variance σ^2, linearly transformed data $X' = aX + b$ have new mean $\mu' = a\mu$ and new variance of $\sigma^{2'} = a^2\sigma^2$. The parameters $a_{c_1}^{R_1}$ and $b_{c_1}^{R_1}$, which linearly transform the distinction distribution of Reader 1 $D_{c_1}^{R_1}(\mu_{c_1}^{R_1}, \sigma^2)$ of confidence c_1 to the expert supervisors' distinction distributions $D_{c_1}^{S}(\mu_{c_1}^{S}, \sigma^2)$ can be estimated as:

$$a_{c_1}^{R_1} = \sqrt{\frac{\sigma^2}{\sigma^2}}, \qquad b_{c_1}^{R_1} = \mu_{c_1}^{S} - (a_{c_1}^{R_1} \times \mu_{c_1}^{R_1}) \tag{2}$$

This transformation results in all readers having a similar subjective quality as the expert supervisor. Thus allowing the conversion of ordinal to interval confidences and the application of statistical evaluations.

2.3.2.2. Reader evaluation metrics. Computation of reader reproducibility requires the evaluation set to be annotated twice by each reader. If A^{1st} and A^{2nd} represent annotated cysts with transformed confidences in the first and second run of annotations, then the similarity between A^{1st} and A^{2nd} representing reader reproducibility can be computed using the Sørensen–Dice index [25] as follows:

$$\text{Reproducibility} = \frac{2 \times \sum_{i=1}^{M}\sum_{j=1}^{N_i} A_{(i,j)}^{1st} A_{(i,j)}^{2nd}}{\sum_{i=1}^{M}\sum_{j=1}^{N_i} A_{(i,j)}^{1st\,2} + \sum_{i=1}^{M}\sum_{j=1}^{N_i} A_{(i,j)}^{2nd\,2}} \tag{3}$$

where $A_{(i,j)}^{1st}$ is the confidence of the jth voxel in the first annotation of the ith B-scan. M and N_i are the number of B-scans and the number of voxels in B-scan i, respectively.

The average of the first and second annotations indicates the final annotation result for each reader for the evaluation set. However, readers will perform with a better reproducibility score when annotating high quality cysts only, therefore reproducibility alone cannot adequately represent the performance of a reader. Thus reader accuracy can be computed as the similarity between the final reader annotation and the reference supervisor annotation using Eq. (3).

2.4. Annotation combination

Finally, upon B-scan annotation by multiple readers, the final annotation can be combined with respect to their accuracy for different cyst confidences, thus for voxel z the final annotation confidence:

$$A^{comb}(v_z) = \frac{\sum_{j=1}^{M} w^{R_j}(v_z) \times A^{R_j}(v_z)}{\sum_{j=1}^{M} w^{R_j}(v_z)} \tag{4}$$

where $A^{R_j}(v_z)$ indicates the confidence of the zth voxel in the annotation of the jth reader. $w^{R_j}(v_z)$ indicates accuracy of jth reader in annotating voxel z, and M is maximum number of readers. $A_{c_i}^{comb}$ represents the combined annotations for voxel z using annotations of M readers. The performance (accuracy) of readers may varies on data with different qualities or data from different modalities. These reader's performance changes are taken in to account in Eq. (4) by weighting the annotation combination based on reader's performance however if the dataset is small or data have similar quality then readers may present similar performance and then weighted annotation combination by Eq. (4) present similar result as simple annotation averaging. In our experiment, the performance of readers in evaluated on cyst Evaluation set and then Eq. (4) is used to combine annotation of different readers.

3. Results

In our experiments, the performance of three readers was evaluated and compared with an expert supervisor based on the cyst evaluation set. The cyst distinction score is leveraged to quantify subjective difference between readers and expert supervisor by modeling the underlying intervals of ordinal cyst confidences for low and high confidence cysts. This is demonstrated in Fig. 5 showing the original Spectralis OCT B-scan (Fig. 5a) and two annotated cysts with the inner region and outer boundary region of the cysts highlighted in red and blue, respectively (Fig. 5b). Cyst 1 is easily identified due to high distinction (0.87) meaning that the histograms of the inner region (red distribution in Fig. 5c) and outer boundary region (blue distribution in Fig. 5c) have little overlap as shown in Fig. 5c. Whereas cyst 2 is difficult to identify due to

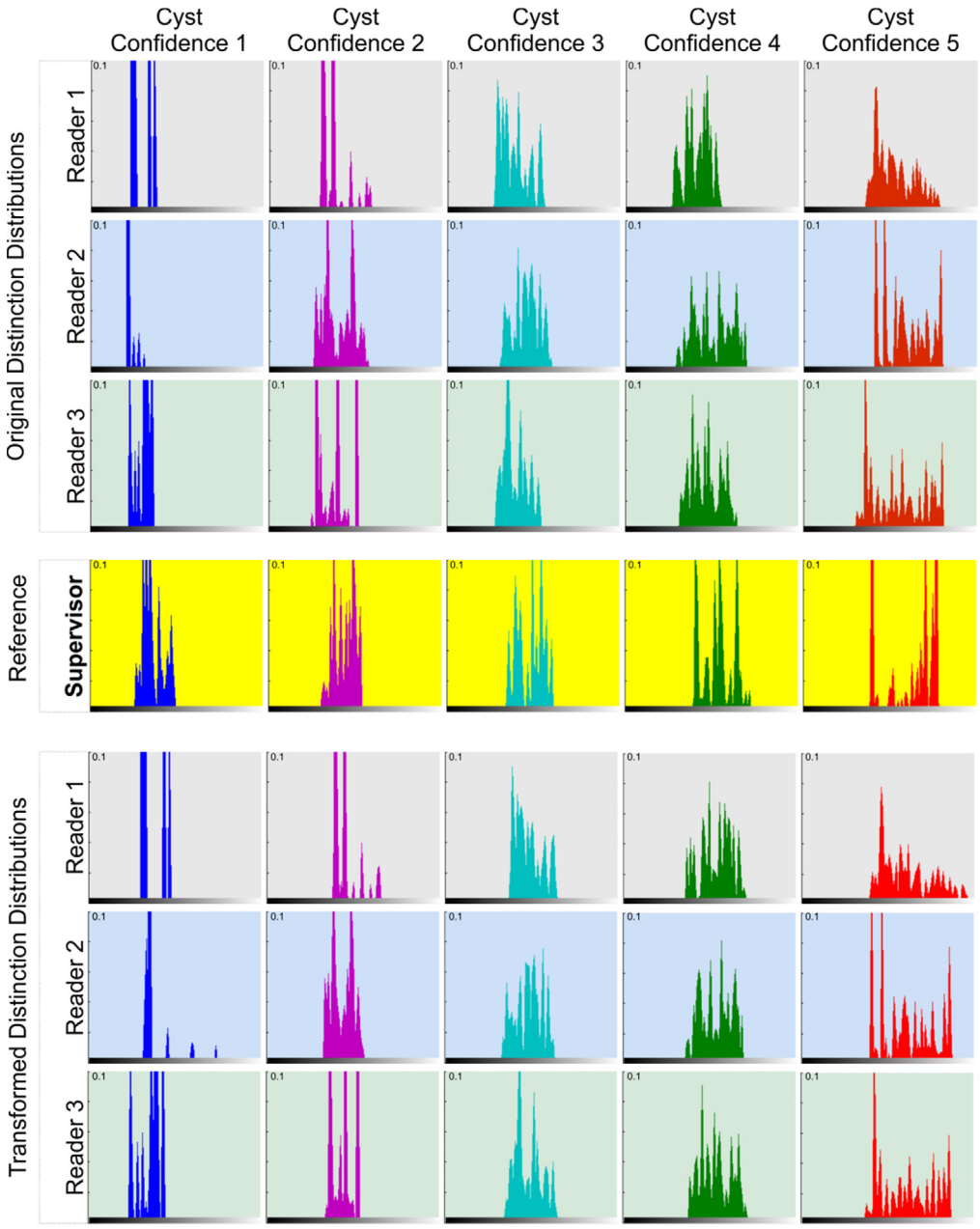


Fig. 6 – Distinction distributions of different readers and one expert supervisor for different cyst confidences before and after transformation. The reference distinction distributions of the expert supervisor are highlighted in yellow. (For interpretation of the references to color in figure legend, the reader is referred to the web version of the article.)

low distinction (0.54) (the histogram of the inner region highlighted in red color highly overlapped with the histogram of the outer boundary region highlighted in blue color as shown in Fig. 5d).

The distribution of cyst distinction scores for different cyst confidences is computed for the 3 readers and the expert supervisor (first 4 rows of Fig. 6). As can be seen, the distinction distributions for low confidence cysts have small distinction

Table 2 – Parameters of distinction distributions.

	Confidence 1 (μ, σ^2)	Confidence 2 (μ, σ^2)	Confidence 3 (μ, σ^2)	Confidence 4 (μ, σ^2)	Confidence 5 (μ, σ^2)
Reader 1	(0.28, 0.05)	(0.35, 0.06)	(0.39, 0.08)	(0.41, 0.07)	(0.53, 0.11)
Reader 2	(0.23, 0.01)	(0.4, 0.09)	(0.46, 0.07)	(0.51, 0.11)	(0.59, 0.12)
Reader 3	(0.31, 0.04)	(0.38, 0.09)	(0.4, 0.07)	(0.47, 0.09)	(0.55, 0.15)
Supervisor	(0.37, 0.06)	(0.44, 0.06)	(0.49, 0.08)	(0.54, 0.09)	(0.62, 0.15)

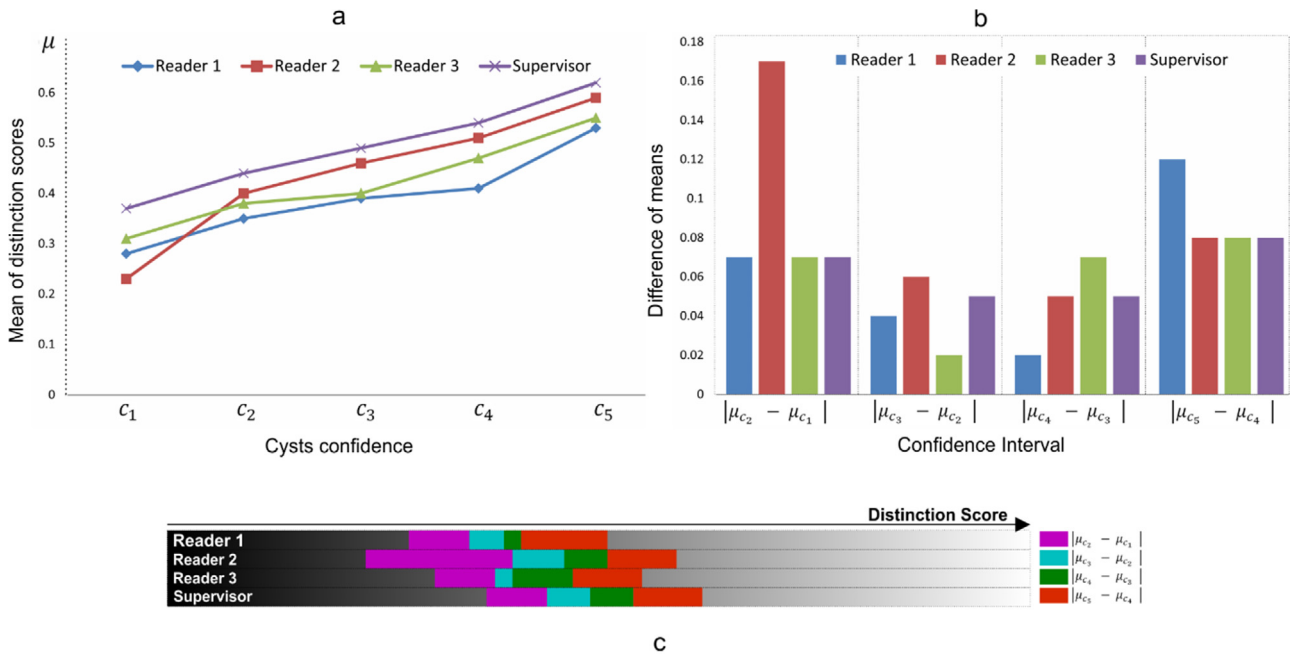


Fig. 7 – (a) Plot of mean distinction distributions of all readers and the expert supervisor for all cyst confidences. (b) Confidence intervals for each reader and the expert supervisor. (c) Confidence intervals for each reader and the expert supervisor with respect to distinction scores. (For interpretation of the references to color near the citation of this figure, the reader is referred to the web version of the article.)

scores and fewer samples in comparison to the distinction distributions of high confidence cysts with high distinction scores. This is due to the difficulty for readers to identify cysts with low distinction. The distinction distributions, the mean and variance of the distributions are presented in Table 2.

The mean of the distinction distributions of all readers and the expert supervisor for different cyst confidences are plotted in Fig. 7a showing the pattern of distinction scores increasing from cysts with confidence 1–5 (c_1–c_5). Cyst confidences feature different intervals for different readers which can be interpreted as the reader subjectivity. The interval between cyst confidences (Fig. 7b) can be computed as the difference between the means of their distinction distributions. Therefore, the cyst confidence interval between cyst confidence 1 and 2 can be computed as $|\mu_{c_2} - \mu_{c_1}|$, in which μ_{c_1} and μ_{c_2} are means of distinction distribution of cyst confidence 1 and 2, respectively. The cyst confidence intervals between cyst confidences 1, 2 and 2, 3 and 3, 4 and 4, 5 are shown in Fig. 7b for the readers and the supervisor. Reader 2 presents the largest cyst confidence interval $|\mu_{c_2} - \mu_{c_1}|$, which is greater than the sum of cyst confidence intervals $|\mu_{c_2} - \mu_{c_1}| + |\mu_{c_3} - \mu_{c_2}|$ of reader 1. This means that most likely some of the annotated cysts by reader 1 with confidence 2 are annotated as a cyst with confidence 1 by reader 2. In order to show how different the readers and the expert supervisor are in terms of subjectivity, the cyst confidence intervals are plotted with respect to the distinction scores as shown in Fig. 7c. The comparison of cyst confidence intervals of reader 1 and the supervisor for cyst confidences 4 and 5 ($|\mu_{c_5} - \mu_{c_4}|$) indicate how they are subjectively different in terms of annotating high confidence cysts as shown in red color in Fig. 7c.

Reader distinction distributions are linearly transformed with respect to the reference distinction distribution to ensure that readers and supervisor have an equivalent subjective quality. The transformation parameters for all readers and all confidences are estimated using Eq. (2) and shown in Table 3. The transformed distinction distributions of each reader can be seen in the last three rows of Fig. 6. The transformed distributions have the same mean and variance as the reference distributions of the expert supervisor. Once readers and expert supervisor have similar subjectivity, statistical evaluations are performed by converting ordinal to interval confidences.

A reader annotation example of a Topcon scan is shown in Fig. 8. As can be seen, the first and second annotations of Reader 3 differ greater in comparison to Readers 1 and 2. Reader 2 presented the highest reproducibility score when their first and second annotations are compared.

Table 3 – Transformation parameters for each reader for different confidences.

	(a_{c1}, b_{c1})	(a_{c2}, b_{c2})	(a_{c3}, b_{c3})	(a_{c4}, b_{c4})	(a_{c5}, b_{c5})
Reader 1	(1.184, 0.038)	(0.954, 0.106)	(0.938, 0.124)	(1.237, 0.033)	(1.35, −0.096)
Reader 2	(4.335, −0.627)	(0.717, 0.153)	(1.034, 0.014)	(0.817, 0.123)	(1.218, −0.098)
Reader 3	(1.464, −0.084)	(0.708, 0.171)	(1.106, 0.047)	(1.042, 0.05)	(0.976, 0.083)

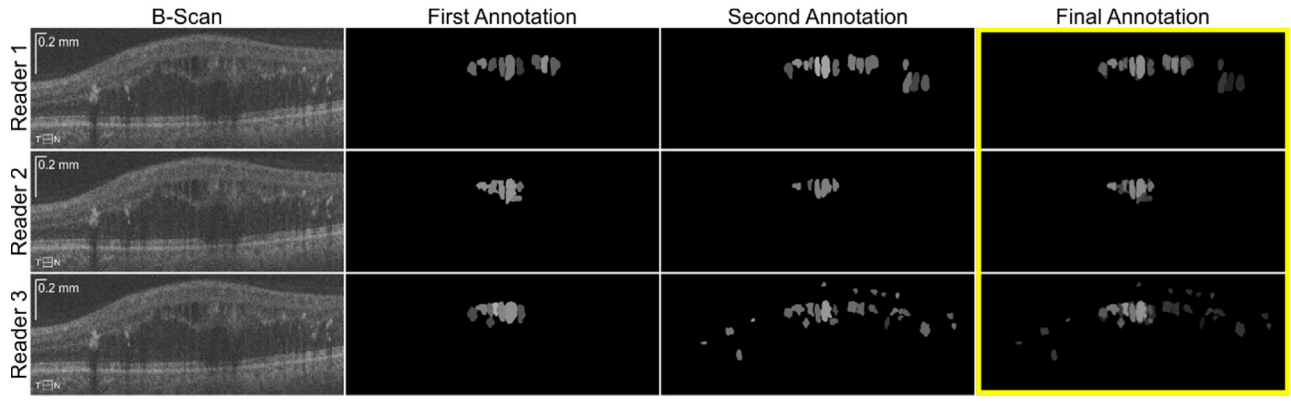


Fig. 8 – Exemplar annotations for three readers showing on the left the original B-scan, followed by the first and second run of annotations and the final annotation on the right outlined in yellow.

Table 4 – Overall accuracy and accuracy per cyst confidence for each reader1.

	Accuracy of readers for cyst confidences					Overall accuracy
	1	2	3	4	5	
Reader 1	0.42	0.78	0.87	0.85	0.83	0.82
Reader 2	0.4	0.71	0.86	0.86	0.85	0.81
Reader 3	0.36	0.76	0.87	0.89	0.89	0.82

Reader accuracy is presented in Table 4. It can be seen that overall reader accuracy is similar but accuracy per confidence varies considerably. For confidence 1, reader accuracy is low due to the difficulty in identifying cysts with low distinction. The readers subjectively also believe that they would not be able to reproduce cysts with confidence 1 themselves. As expected, accuracy increases with cyst distinction, with the highest accuracy seen at confidence 4 and 5 by Reader 3 with a value of 0.89 and the highest accuracy for cysts with a confidence 1 is achieved by Reader 1 with a value of 0.42.

After reader performance evaluation, their annotations are combined with respect to their accuracy for different cyst confidences using Eq. (4). The effectiveness of multiple annotations by complementary readers is revealed by comparing the combined annotations of different readers as shown in Figs. 9–12. The original B-scan from a Cirrus OCT scan and its reference annotation by the expert supervisor are shown in Fig. 9a and e. The annotations of Reader 1–3 are shown in Fig. 9b–d. Reader 2 failed to annotate the cysts properly as shown in Fig. 9c while the cysts are successfully annotated by Readers 1 and 3 (Fig. 9b and d). Due to multiple annotations, the missing cyst by Reader 2 is detected in combination with Readers 1 and 3 as shown in Fig. 9f and h.

The confidence based annotation helps readers to identify more cysts even when they are difficult to detect as shown in Fig. 10. The combined annotations indicate more low confidence cysts (Fig. 10f–h) in comparison to the reference annotation (Fig. 10e). Further examples from Spectralis and Nidek scans can be seen in Figs. 11 and 12, respectively. In all examples, the combined results have high compatibility with the reference annotations, showing how complementary readers generate more accurate results.

Fig. 13 presents 1 example scan from each vendor demonstrating the partitioned subsets used to build the benchmark data set for cyst segmentation, where partition 1 is annotated by Readers 1 and 2, partition 2 by Readers 2 and 3, and partition

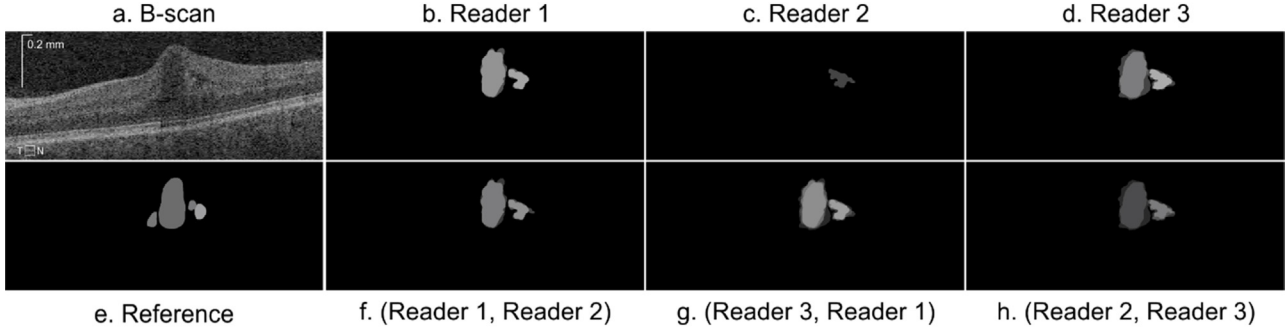


Fig. 9 – Example of combined annotations from a Cirrus scan. (a) Original B-scan, (b–d) annotation of Readers 1–3. (e) Reference annotation. (f–h) Combined annotations from annotations of Readers 1 and 2, Readers 3 and 1, and Readers 2 and 3, respectively.

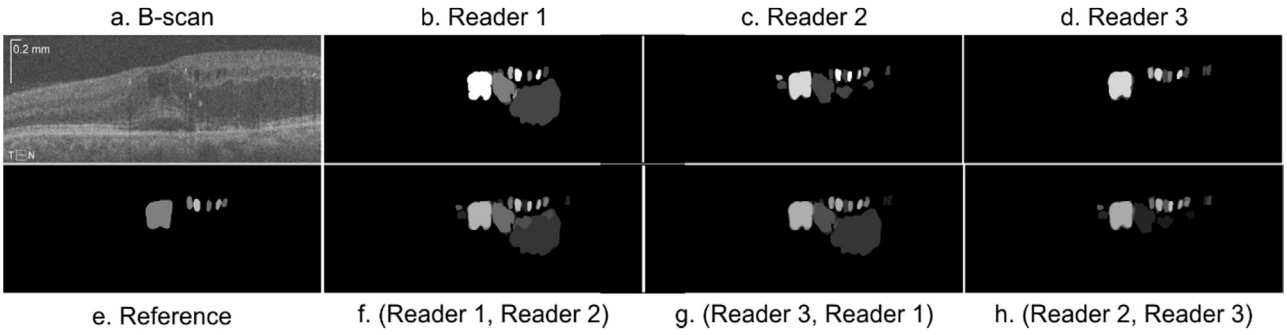


Fig. 10 – Example of combined annotations from a Topcon scan. (a) Original B-scan, (b–d) annotation of Readers 1–3. (e) Reference annotation. (f–h) Combined annotations from annotations of Readers 1 and 2, Readers 3 and 1, and Readers 2 and 3, respectively.

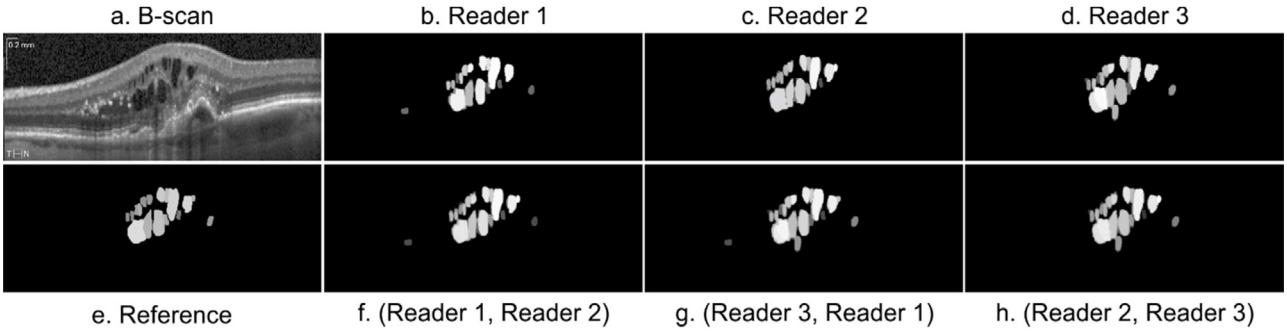


Fig. 11 – Example of combined annotations from a Spectralis scan. (a) Original B-scan, (b–d) annotation of Readers 1–3. (e) Reference annotation. (f–h) Combined annotations from annotations of Readers 1 and 2, Readers 3 and 1, and Readers 2 and 3, respectively.

3 by Readers 1 and 3. As can be seen, multiple annotations by two different readers are combined with respect to their performance on the cyst evaluation set to generate the final annotation (Fig. 13d).

4. Discussion

Cyst detection is of high clinical relevance because cysts could be identified as an imaging biomarker from OCT. It is shown in earlier studies that cysts are the only prognostic parameter for functional outcome in neovascular age-related macular degeneration (nAMD) therapy. Ritter et al. [7] have shown from data of a large multi-center phase III clinical trial of patients with nAMD (MONTBLANC) that all patients with intraretinal cysts at baseline have shown a significantly lower mean best corrected visual acuity (BCVA) than patients without cysts at baseline. This is true for any morphologic combination of other morphologic changes such as subretinal fluid or pigment epithelial detachment at baseline together with cysts.

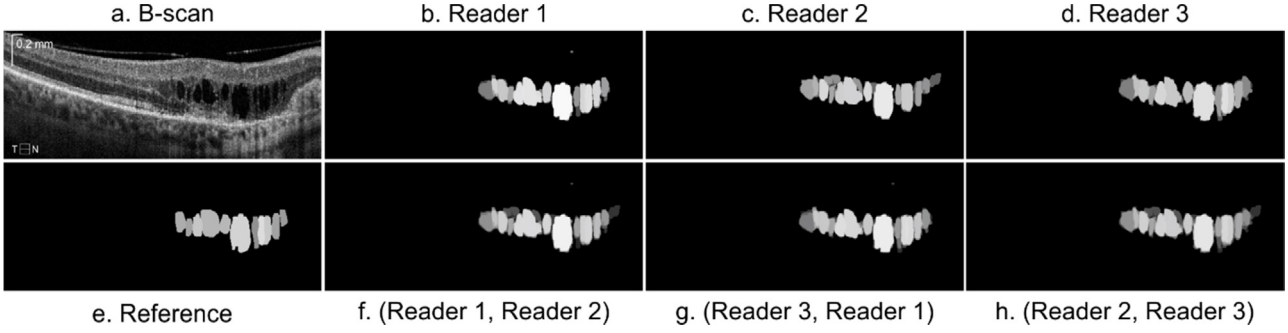


Fig. 12 – Example of combined annotations from a Nidek scan. (a) Original B-scan, (b–d) annotation of Readers 1–3. (e) Reference annotation. (f–h) Combined annotations from annotations of Readers 1 and 2, Readers 3 and 1, and Readers 2 and 3, respectively.

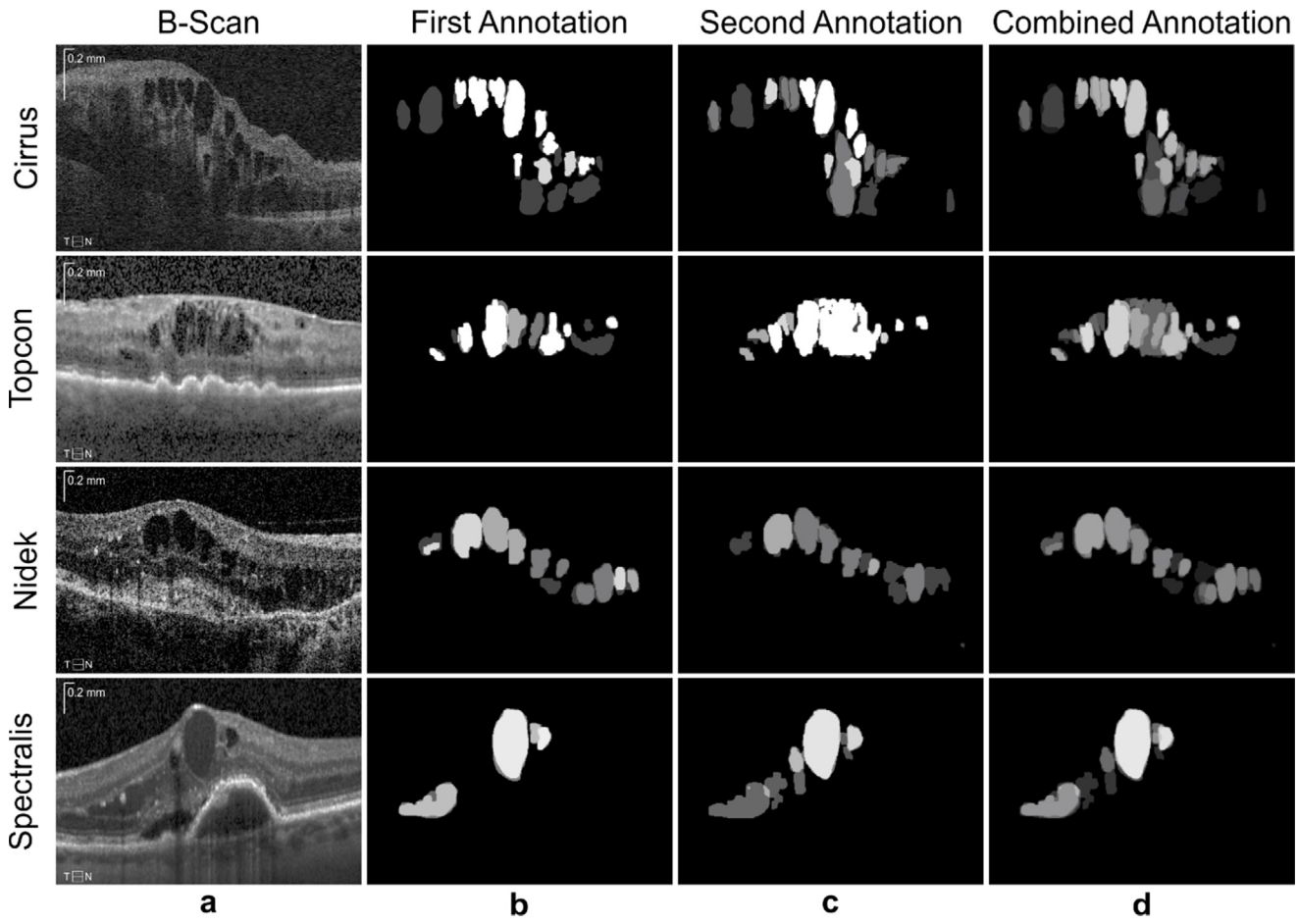


Fig. 13 – Annotated B-scan from all four vendors from the benchmark dataset. (a) Original B-scans, (b) annotation of Reader 1. (c) Annotation of Reader 2. (d) Combined annotation from annotations of Readers 1 and 2.

No other combination of morphologic changes at baseline had a functional outcome as bad as cysts [7]. These findings could also be seen from other studies on prognostic morphologic factors for BCVA in nAMD such as the EXCITE study [6] and the CATT study [26], as well as from the RESTORE study on patients with diabetic macular edema. Here Gerendas et al. have shown that not only the appearance of cysts but also the size of the cysts can affect functional outcome. In particular, cysts with an overall height of more than 380 μm at baseline have shown a significantly higher BCVA at baseline and maintained better BCVA until the end of the study. Patients with large cysts had the same BCVA gain as patients with small cysts but the gap between both groups remained stable (Gerendas BS, IOVS, 2014, ARVO E-Abstract 228). These findings are only based on time-domain OCT scans with 6 radial cuts through the retina where manual evaluation is feasible and easy to manage. However for SD-OCT with many hundreds of B-scans to evaluate, other methods to detect these important biomarkers are needed. This requirement has led to the work presented here of a method to structure manual annotation and evaluate (semi-)automatic algorithms for cyst detection.

An important contribution of this work is the proposal of a benchmark framework for ophthalmic SD-OCT and generating a confidence based benchmark dataset for cyst segmentation that can be used to reliably evaluate cyst segmentation methods. The generated benchmark dataset aims to resolve aforementioned limitations by firstly incorporating scans from the 4 major SD-OCT scanner vendors, allowing the inclusion of vendor specific image variations, thus resulting in a highly diverse and robust training and testing set. In addition, scan composition ranges from 5 to 256 B-scans as well as featuring varying image qualities, thus ensuring an accurate "real world" scan representation.

Conventional retinal image data sets [11,14,16,27,28] are annotated in a binary fashion which is not ideal for cyst annotation due to the difficulty in delineating cysts from non-cysts due to noise and other imaging artifacts. This may result in low confidence cysts being missed completely, affecting the performance of machine learning based methods. Our proposed confidence based annotation protocol tackles this problem by assigning cyst confidences to segmented cyst regions with respect to their shape, distinction, continuity, and position. The confidences let readers identify more cysts even when they difficult to delineate. Despite confidence based annotation being more time consuming in comparison to binary annotation, it is necessary and important as it provides a better understanding of cyst appearance and cyst volume

which may affect patient vision. Moreover machine learning methods can learn more from confidence based annotations and it is straightforward to convert them to binary if required.

The cost of annotation in terms of human resources and time pushes researchers to use a single reader to annotate a limited number of data, increasing bias toward reader and data. The proposed benchmark is generated from multiple annotations of complementary readers to remove this bias. In multiple annotations, if one of readers fails to annotate a particular cyst, the chance still remains to detect it by another reader's annotation. Moreover data evaluation processes are applied to split a dataset into subsets of data mimicking the original dataset, removing annotation bias toward particular data or vendors and to save annotation time. Finally an intelligent strategy is used to evaluate reader performance and combine their annotations with respect to reader performance to generate the final annotations. The final confidence based cyst annotations from 26 SD-OCT volume scans comprise the presented benchmark database.

To the best of our knowledge, this is the first study which generates a comprehensive and confidence based benchmark dataset for the evaluation of (semi-)automatic cyst segmentation algorithms. The benchmark database can be made publicly available to encourage researchers to develop new cyst segmentation algorithms. Future work could concentrate on establishing perceptual measures such as cyst distortion and cyst detection ratio for validating cyst segmentation methods. The proposed methods for creating benchmark datasets for retinal SD-OCT annotation can be used to evaluate cyst segmentation algorithms as well as algorithms for the detection of subretinal fluid, sub-RPE fluid or any other segmentable structures on the images. This benchmark dataset can serve as a dataset of cysts with high specificity whereas other annotation definitions might lead to less conservative ground truth annotations. The exact definition of what should be segmented must always be made beforehand and will improve both the quality of the benchmark dataset as well as the quality of any segmentation algorithm.

REFERENCES

[1] M. Niemeijer, B. Van Ginneken, M.J. Cree, et al., Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs, IEEE Trans. Med. Imaging 29 (1) (2010) 185–195.

[2] G. Quellec, K. Lee, M. Dolejsi, et al., Three-dimensional analysis of retinal layer texture: identification of fluid-filled regions in SD-OCT of the macula, IEEE Trans. Med. Imaging 29 (6) (2010) 1321–1330.

[3] M.D. Abramoff, M.K. Garvin, M. Sonka, Retinal imaging and image analysis, IEEE Trans. Med. Imaging 3 (2010) 169–208.

[4] R.A. Costa, M. Skaf, L.A. Melo Jr., et al., Retinal assessment using optical coherence tomography, Prog. Retin. Eye Res. 25 (3) (2006) 325–353.

[5] M. Baroni, P. Fortunato, A. La Torre, Towards quantitative analysis of retinal features in optical coherence tomography, Med. Eng. Phys. 29 (4) (2007) 432–441.

[6] C. Simader, M. Ritter, M. Bolz, et al., Morphologic parameters relevant for visual outcome during anti-angiogenic therapy of neovascular age-related macular degeneration, Ophthalmology 121 (6) (2014) 1237–1245.

[7] M. Ritter, C. Simader, M. Bolz, et al., Intraretinal cysts are the most relevant prognostic biomarker in neovascular age-related macular degeneration independent of the therapeutic strategy, Br. J. Ophthalmol. 98 (12) (2014) 1629–1635.

[8] A.H. Kashani, P.A. Keane, L. Dustin, et al., Quantitative subanalysis of cystoid spaces and outer nuclear layer using optical coherence tomography in age-related macular degeneration, Investig. Ophthalmol. Vis. Sci. 50 (7) (2009) 3366–3373.

[9] S. Roychowdhury, D.D. Koozekanani, et al., Automated localization of cysts in diabetic macular edema using optical coherence tomography images, in: 35thAnnual International IEEE Conference on Engineering in Medicine and Biology Society (EMBC), 2013.

[10] D. Huang, E.A. Swanson, C.P. Lin, et al., Optical coherence tomography, Sci. Mag. 254 (5035) (1991) 1178–1181.

[11] A.F. Fercher, C.K. Hitzenberger, W. Drexler, et al., In vivo optical coherence tomography, Am. J. Ophthalmol. 116 (1) (1993) 113–114.

[12] W. Geitzenauer, C.K. Hitzenberger, U.M. Schmidt-Erfurth, Retinal optical coherence tomography: past, present and future perspectives, Br. J. Ophthalmol. 95 (2) (2010) 171–177.

[13] J. Staal, M.D. Abràmoff, M. Niemeijer, et al., Ridge-based vessel segmentation in color images of the retina, IEEE Trans. Med. Imaging 23 (4) (2004) 501–509.

[14] B. Al-Diri, A. Hunter, D. Steel, et al., REVIEW – a reference data set for retinal vessel profiles, Eng. Med. Biol. Soc. 30 (2008) 2262–2265.

[15] A. Hoover, V. Kouznetsova, M. Goldbaum, Locating blood vessels in retinal images by piece-wise threshold probing of a matched filter response, IEEE Trans. Med. Imaging 19 (3) (2000) 203–210.

[16] H.A. Kirisli, M. Schaap, C.T. Metz, et al., Standardized evaluation framework for evaluating coronary artery stenosis detection, stenosis quantification and lumen segmentation algorithms in Computed Tomography Angiography, Med. Image Anal. 17 (8) (2013) 859–876.

[17] G. Wilkins, O. Houghton, A. Oldenburg, Automated segmentation of intraretinal cystoid fluid in optical coherence tomography, IEEE Trans. Biomed. Eng. 59 (4) (2012) 1109–1114.

[18] H.R. Marateb, M. Mansourian, P. Adibi, et al., Manipulating measurement scales in medical statistical analysis and data mining: a review of methodologies, J. Res. Med. Sci. 19 (1) (2014) 47–56.

[19] P. Nagpaul, Guide to Advanced Data Analysis using IDAMS Software, New Delhi, India, 2001.

[20] E. Frank, H. Mark, A Simple Approach to Ordinal Classification, Springer, Berlin/Heidelberg, 2001.

[21] S. Labovitz, The assignment of numbers to rank order categories, Am. Sociol. Rev. 35 (1970) 515–524.

[22] L. Mayer, A note on treating ordinal data as interval data, Am. Sociol. Rev. 36 (1971) 519–520.

[23] M.P. Allen, Conventional and optimal interval scores for ordinal variables, Sociol. Methods Res. 4 (1976) 475–494.

[24] J.S. Granberg-Rademacker, An algorithm for converting ordinal scale measurement data to interval/ratio scale, Educ. Psychol. Meas. 70 (2010) 74–90.

[25] T. Sørensen, A method of establishing groups of equal amplitude in plant sociology based on similarity of species

and its application to analyses of the vegetation on Danish commons, in: I kommission hos E. Munksgaard, 1948.

[26] G.J. Jaffe, D.F. Martin, C.A. Toth, et al., Macular morphology and visual acuity in the comparison of age-related macular degeneration treatments trials, Ophthalmology 120 (9) (2013) 1860–1870.

[27] X. Chen, M. Niemeijer, L. Zhang, et al., Three-dimensional segmentation of fluid-associated abnormalities in retinal OCT: probability constrained graph-search-graph-cut, IEEE Trans. Med. Imaging 31 (8) (2012) 1521–1531.

[28] D.C. Fernandez, Delineating fluid-filled region boundaries in optical coherence tomography images of the retina, IEEE Trans. Med. Imaging 24 (8) (2005) 929–945.