PLOS ONE

RESEARCH ARTICLE

Statistical framework for validation without ground truth of choroidal thickness changes detection

Tiziano Ronchetti[1,2,3]*, Christoph Jud[1☯], Peter M. Maloca[3,4,7,8☯], Selim Orgül[4], Alina T. Giger[1], Christoph Meier[2], Hendrik P. N. Scholl[4,8,9], Rachel Ka Man Chun[5], Quan Liu[5,6], Chi-Ho To[5,6], Boris Považay[2‡], Philippe C. Cattin[1‡]

1 Department of Biomedical Engineering (DBE), University of Basel, Basel, Switzerland, **2** Institute for Human Centered Engineering (HuCE)-optoLab, Bern University of Applied Sciences, Bern, Switzerland, **3** OCTlab, Department of Ophthalmology, University Hospital Basel, Basel, Switzerland, **4** Department of Ophthalmology, University of Basel, Basel, Switzerland, **5** School of Optometry, The Hong Kong Polytechnic University (PolyU), Hong Kong, PR China, **6** State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, PR China, **7** Moorfields Eye Hospital, London, United Kingdom, **8** Institute of Molecular and Clinical Ophthalmology Basel (IOB), Basel, Switzerland, **9** Wilmer Eye Institute, Johns Hopkins University, Baltimore, Maryland, United States of America

☯ These authors contributed equally to this work.
‡ These authors also contributed equally to this work.
* Tiziano.Ronchetti@unibas.ch

Citation: Ronchetti T, Jud C, Maloca PM, Orgül S, Giger AT, Meier C, et al. (2019) Statistical framework for validation without ground truth of choroidal thickness changes detection. PLoS ONE 14(6): e0218776. https://doi.org/10.1371/journal.pone.0218776

Editor: Sanjoy Bhattacharya, Bascom Palmer Eye Institute, UNITED STATES

Received: September 30, 2018

Accepted: June 11, 2019

Published: June 28, 2019

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: This study was supported by the Swiss National Science Foundation within the SNSF Project 320030 146021 "Characterization of choroidal changes in children and its temporal response to optical defocus." This work is a significant part of an international cooperation with clinical partners in Guangzhou and Hong Kong, supported by the National Natural Science

Abstract

Monitoring subtle choroidal thickness changes in the human eye delivers insight into the pathogenesis of various ocular diseases such as myopia and helps planning their treatment. However, a thorough evaluation of detection-performance is challenging as a ground truth for comparison is not available. Alternatively, an artificial ground truth can be generated by averaging the manual expert segmentations. This makes the ground truth very sensitive to ambiguities due to different interpretations by the experts. In order to circumvent this limitation, we present a novel validation approach that operates independently from a ground truth and is uniquely based on the common agreement between algorithm and experts. Utilizing an appropriate index, we compare the joint agreement of several raters with the algorithm and validate it against manual expert segmentation. To illustrate this, we conduct an observational study and evaluate the results obtained using our previously published registration-based method. In addition, we present an adapted state-of-the-art evaluation method, where a paired t-test is carried out after leaving out the results of one expert at the time. Automated and manual detection were performed on a dataset of 90 OCT 3D-volume stack pairs of healthy subjects between 8 and 18 years of age from Asian urban regions with a high prevalence of myopia.

Introduction

The choroid is a vascular structure located at the posterior part of the uveal tract in the eye, between the relative rigid sclera and the more flexible, light-sensitive retina. The choroid plays

Foundation of China NSFC, both PR China and numerous institutions in Switzerland (Basel, Bern, Luzern and Biel).

Competing interests: The authors have declared that no competing interests exist.

a crucial role in the ocular metabolism circulation providing oxygen and metabolites to the outer retina [1, 2]. Its variable thickness depends on factors such as blood pressure, axial length and age [3].

While in adults it has consistently been assessed that the choroidal thickness decreases with advancing age [4–6], studies researching choroidal development during childhood and adolescence led to contradictory conclusions. Subfoveal choroidal thickness was found to be positively correlated with age in Caucasian [7–10], but negatively in Asian children, where the prevalence of myopia is significantly higher [5, 11]. The choroid plays an active role in emmetropization, both by modulation of its thickness to adjust the retina to the optical focus plane (choroidal accommodation) and through the regulation of the scleral growth [12, 13]. Its complex interaction with other tissues as well as its strong dependence on many other factors like blood pressure or diurnal variations, demand a precise and reliable monitoring method [1, 14].

Longitudinal studies of teenagers who develop myopia documented an eye ball elongation. This process is associated with significant thinning of the choroidal thickness in cases of high myopia [15]. Therefore, choroidal thickness, but also choroidal structure, is considered to be an important marker for monitoring myopic progression and for predicting myopia. The main challenge for detecting disease progress is to recognize particular minute changes as early as possible.

Based on optical coherence tomography (OCT) imaging [16] that unveils highly resolved details of the retina and choroid, there are segmentation- and registration-based methods for the detection of temporal changes in the thickness of the choroid (a detailed description of the different state-of-the-art methods follows below). In most cases the evaluation of such methods is performed by an artificially generated ground truth (usually the average of expert segmentations). A major drawback of such an approach is that differences between the manual segmentations cannot be correctly taken into account, when equal weights are given to all expert segmentations during averaging.

In this paper, our primary motivation is to show how to evaluate the performance of a method for choroidal thickness changes detection without generating an artificial average ground truth. We present a validation framework, purely based on common agreement, to assess the detection-performance, using the Williams' Index [17] as measure. As example of an automatic detection method we use our recently proposed registration-based method CRAR [18]. Additionally, we present an adapted state-of-the-art approach, where an artificial ground truth is created by averaging the results of the remaining experts after leaving one out at the time. In an observational study we examined long-term changes in the choroidal thickness of 90 OCT 3D-volume stack pairs of Chinese subjects between 8 and 18 years of age. For each eye, measurements were collected twice within a period of at least 3 to maximum 14 months.

Our paper's contributions are: 1) We present a statistical validation framework for automated choroidal thickness changes detection applicable in cases where a real ground truth is not available. 2) We demonstrate the framework's reliability by evaluating the results of our registration-based algorithm CRAR against those obtained by the experts. 3) We extend the commonly used power analysis approach by leave-one-out cross validation to become an ideal component of our statistical framework. 4) Based on a clinical study with volunteers with a high prevalence of juvenile myopia, we gain insight into possible correlations between time interval (between the measurements) and the choroidal thickness changes measured.

To the best of our knowledge, this is the first time that a statistical validation framework for automated choroidal thickness changes detection combines a method purely based on common agreement with an exhaustive power analysis approach.

Background and prior work

OCT has become the main contactless, non-invasive method to characterize changes in the corneal, retinal and choroidal structures and to monitor eye growth [16]. Operating in the near infrared range, OCT provides high resolution imaging within a micrometer range and is well established in ophthalmology. Longitudinal studies using OCT imaging offer a unique possibility to gain insight into the dynamics of anatomical changes in the retina and choroid. This leads to an understanding of the mechanisms regulating such processes.

The commonly used representation of the thickness of the choroid is the choroidal thickness map, see Fig 1(a), generated from the pre-segmented data of a 3D-volume stack (C-scans) consisting of adjacent sagittal tomograms (B-scans). For every depth scan (A-scan) the difference between two segmentation-planes at this lateral location is calculated and represented. The choroidal thickness is defined as the vertical distance (along the A-scans/in z-direction) between the Bruch's Membrane (BM) and the Choroid-Sclera Interface (CSI), see Fig 1(b). Segmentation is frequently performed as a manual task to determine the ground truth from OCT-measurements. Due to the lack of alternative high precision in-vivo measurement methodologies, signals are typically compared to their histological equivalent. This proves to be notoriously difficult, subjective, and unreliable in view of the large amount of data points and the weak signals that are frequently hard to interpret for the human observer [14]. Even the intra-observer reproducibility is relatively low and with novel algorithms that excel in this regard a different approach to verify the reliability has to be taken. Automated detection of noisy and speckled OCT-images at the low-signal-end of the depth scan already has a long history and usually focuses on segmentation by delineating borders that are associated with large scale changes of the refractive index, or by determining tissue texture appearance.

Among the current approaches for detecting choroidal boundaries, graph search based segmentation methods represent a state-of-the-art [19–21]. However, their performance is limited by the low contrast of the choroidal boundaries, the inhomogeneity of the choroid's texture and great variation of the choroidal thickness [2]. Recently, a segmentation algorithm was presented, which combines a robust contour-detection method with a graph search, based on a novel weighting scheme [22]. However, the reliability of this method depends strongly on the choice of nodes and weights. In [23], CSI segmentation was performed using an improved graph search algorithm with curve smooth constraints. However, this approach was especially

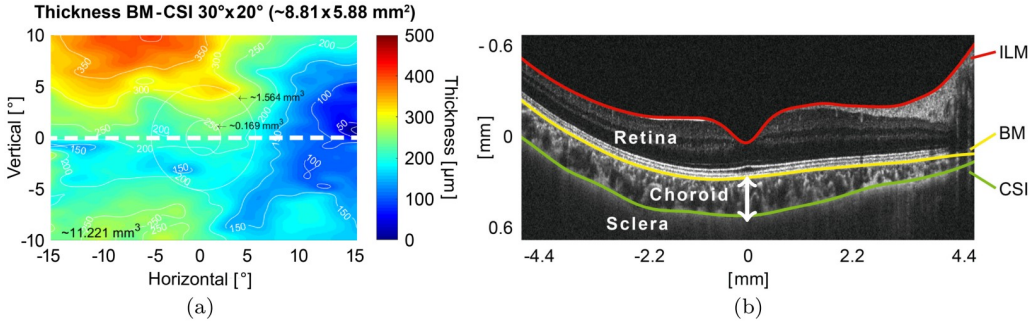

Fig 1. Choroidal thickness map and OCT B-scan with segmented layers. (a) Visualization of the choroidal thickness (BM-CSI) including the choroid's measured volume of a healthy right eye based on graph search algorithm. Circles indicate the location of the macula. The BM-CSI volume of the whole C-scan is indicated in the bottom left. (b) B-scan, a sagittal cross-section of the posterior eye segment through the retina, choroid and sclera, separated by the layers ILM, BM and CSI (source: Hydra, HuCE-optoLab/BUAS). The image was cut off in the vertical/z-direction for better visualization. The full A-scan length is 1.9 mm. The ILM is the Inner Limiting Membrane, while BM and CSI denote the Bruch's Membrane and Choroid-Sclera Interface, respectively.

https://doi.org/10.1371/journal.pone.0218776.g001

developed for Cirrus HD-OCT (High Definition) and still needs to be extended to other OCT devices. The use of a convolutional network architecture, where an optimal graph-edge weight can be learned directly from raw pixels, was proposed in [24]. However, this approach requires a huge amount of training data (approximately 1000 manually segmented B-scans) from the experts. In [25] the authors presented a method for unsupervised learning to identify anomalies in imaging data as candidates for pathological markers. However, it still has to be proven that the method is really able to recognize very small changes, for example in an early stage of a disease. Despite progress in image processing the use of single frame segmentation is inherently difficult, especially in longitudinal clinical studies where successive imaging sessions can strongly vary in signal quality.

In order to circumvent this limitation, we recently proposed CRAR, a method to detect early **C**horoidal changes using piecewise rigid image **R**egistration and eye-shape **A**dherent **R**egularization [18]. Our method is a registration-based approach specifically developed for longitudinal studies, allowing to overcome critical problems like low contrast, loss of signal and the presence of artifacts, which are yet unsolved by most segmentation-based methods for the detection of minute choroidal changes. It needs to be emphasized that the aim of CRAR is not to localize the exact position of the CSI, i.e. the exact cutting line between choroid and sclera, but to figure out the displacement in anterior-posterior/z-direction of its surrounding area within a specific time interval. Based on the resulting distortion fields, the use of a roughly localized CSI borderline is already sufficient to extract the corresponding volume changes.

In a previous paper [18] we already demonstrated CRAR's robustness regarding noise by testing the method on scan-rescans. As by scan-rescan no changes occur except the noise, the detected displacement field must be close to zero. By inducing synthetic deformations in the area of the CSI, we attested CRAR's applicability on follow-ups as well as its ability to detect changes as small as 5 μm in the thickness of the choroid [18]. For more details about CRAR the reader is also referred to the S1 Appendix.

Method

To deal with the lack of valid ground truth information, we compare the results achieved by an automatic technique with those of a group of human experts, see Fig 2. In this context, one assumes that human raters hold some prior knowledge of a "mental" ground truth that is reflected in their manual tracings [26, 27]. Human rater accuracy and variability is taken into account by measuring the similarity between the expert and automatic segmentation [28]. Since a solid ground truth does not exist, the most natural solution is an evaluation based on common agreement.

The key idea is the following: we define a method X to be at level with a group $\{Y_i\}$ of experts if X agrees with each expert Y_i at least as often as the experts $\{Y_i\}$ agree among themselves. In other words, with the help of an agreement index, we show that the agreement rate between algorithm X and each expert is at least as high as within the expert group itself. First, we apply the similarity measures commonly used for evaluations of segmentation results. Since we are not looking for a surface but a contour (the CSI, the lower boundary of the choroid), we need a more appropriate measure for contours. To show how to proceed when working with a registration approach, where the outcome is the displacement of the contour line during the time interval between the two measurements (and not the contour itself), we use our previously published algorithm CRAR as example for method X. In this case, the displacement of the contour line is a very natural and intuitive metric to be used. The algorithm's performance is evaluated with an agreement test by comparing the displacements detected by CRAR with those of the experts. As an agreement measure we opt for the Williams' Index [29]

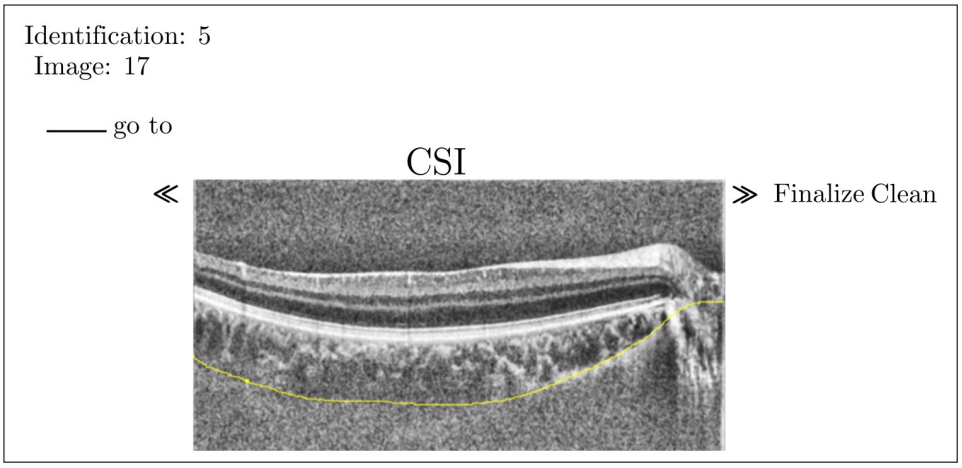

Fig 2. Sample screen of our online tool for manual expert segmentation. According to the consensus between the experts, interconnecting tissues and vessels inside the sclera were ignored while the yellow segmentation line was continued on the side of the optical nerve horizontal. The pre-processing (filtering and histogram equalization) for better contrast during the task was activated in this case by the expert.

https://doi.org/10.1371/journal.pone.0218776.g002

and demonstrate that the algorithm's performance is independent of the chosen metric. In addition, for comparison with a state-of-the-art evaluation procedure, we conduct a paired t-test using an artificially generated ground truth.

Intra-rater coefficient

In order to quantify the reliability of each expert in segmenting the same image we introduce the corresponding Intra-Rater Coefficient. Let us consider a set of S images, namely OCT B-scans of size $m \times n$ pixels. Each image has to be segmented thrice by an expert. Considering the s^{th} image segmented by expert j, the deviation $\Delta\mathcal{E}_j^s = \mathcal{E}_j^s - \bar{\mathcal{E}}_j^s$ of any one of the three segmentations $\mathcal{E}_j^s$ from their average $\bar{\mathcal{E}}_j^s$ is calculated. The number #counts(j, l) of all such deviations, which are within a given tolerance interval $[-l, l]$, i.e.

$$\#\text{counts}(j,l) = \#\{\Delta\mathcal{E}_j^s \mid -l \leq \Delta\mathcal{E}_j^s \leq l\}, \text{ where } s \in \{1,\ldots,S\},$$

leads to the definition of the Intra-Rater Coefficient IRC_j of expert j

$$\text{IRC}_j = \frac{\#\text{counts}(j,l)}{3nS}. \tag{1}$$

Here, $3nS$ denotes the total number of pixels to be segmented by each expert and l is the margin of tolerance set for the manual segmentation (in our case $l = 20$ μm or 5 pixels).

Now, we try to get a feeling for the value of this coefficient. It can be shown that a rater that repeats the segmentation of the CSI at random will achieve an IRC of at most 0.2. Such a rater can be simulated by inducing, in the manually annotated scans, synthetic B-spline deformations with randomly chosen coefficients for the linear combinations of their basis functions. In order to attest reliability to the manual segmentations, we aim to achieve IRC values of at least 0.70. This value corresponds to a variance in the segmentations of one rater of approx. 25 μm.

Williams' index

According to Williams [17] and [29] we propose an agreement index giving an answer to the following question: given a group of $r \geq 3$ raters labeling a finite set of pixels, does rater j agree

with a group of experts in the same way as the group members agree among each other? If this is the case, the index of agreement is set equal to 1.

Let $\mathcal{E}_j$, $\mathcal{E}_{j'}$ and $\mathcal{E}_{j''}$ be the results of all the manual segmentations by the experts j, j' and j'', the Williams' agreement Index (WI) of expert j is defined as:

$$WI_j = \frac{(r-2)\sum_{j'\neq j}^{r} s(\mathcal{E}_j, \mathcal{E}_{j'})}{2\sum_{j'\neq j}^{r}\sum_{j''\neq j}^{j'-1} s(\mathcal{E}_{j'}, \mathcal{E}_{j''})}, \tag{2}$$

where $s(\mathcal{E}_j, \mathcal{E}_{j'}) \in [0, 1]$ provides a quantification for the similarity between the predictions of rater j and j' for all pixels (for more details about $s(\mathcal{E}_j, \mathcal{E}_{j'})$ see the next subsection). The ratio derived is compared to the value of 1. If this index is greater than 1, it can be concluded that rater j agrees with the other raters at least as often as they agree with each other [29]. Otherwise, the rate of agreement obtained between rater j and the group of raters is smaller than the rate of agreement within the group of raters.

Similarity measures

The main underlying principle of our evaluation is the definition of agreement between a method X (here CRAR) for automated detection of choroidal thickness changes and a group of experts doing manual segmentations. The agreement of two experts is defined as the similarity between their respective segmentations. An intuitive similarity measure for the comparison between our algorithm and manual segmentations is the difference between the automatically detected changes and the difference between manual segmentations by experts at different times. The evaluation framework is established by first calculating the WI using the common similarity measures for segmentation, i.e. the Dice Coefficient (DC) and the Jaccard Similarity Coefficient (JC). Such similarity measures are typically applied to survey the segmentation of surfaces.

As our task is to recognize the CSI, the lower contour of the choroid, we need a more appropriate similarity measure. For this reason we opt for the Bidirectional Local Distance (BLD, see Fig 3, for more detail see [30]), a more robust and conclusive similarity measure for surfaces than the DC and JC, as shown in Fig 4.

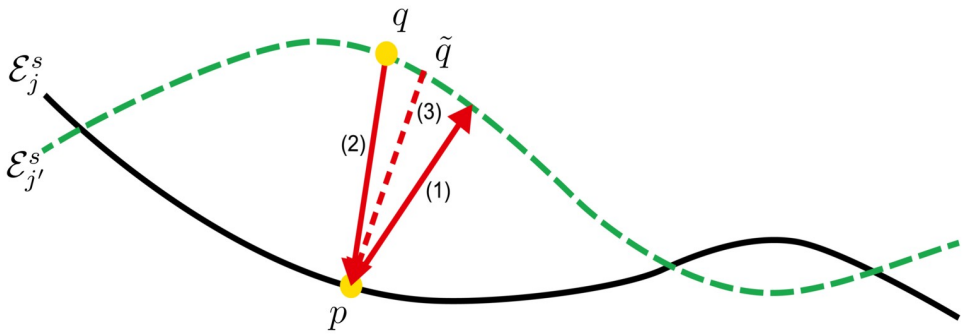

Fig 3. The calculation of the similarity measure BLD. First, the minimum "forward" distance $d_{\min}(p, \mathcal{E}_{j'}^s)$ between the point $p \in \mathcal{E}_j^s$ and the contour $\mathcal{E}_{j'}^s$ is determined, here marked as (1). Second, among all the points q on $\mathcal{E}_{j'}^s$ with a "inverse" minimum distance $d_{\min}^{-1}(q, \mathcal{E}_j^s)$, those are selected whose minimal distance is found at the point p. Here, q and $\tilde{q}$ are the candidates, with the corresponding distances denoted by (2) and (3). Then, the maximum distance among the candidates, in this case (2), is chosen as $d_{\max}^{-1}(\mathcal{E}_{j'}^s, p)$. Finally, $\mathrm{BLD}(p, \mathcal{E}_{j'}^s)$ is defined as the maximum between $d_{\min}(p, \mathcal{E}_{j'}^s)$ and $d_{\max}^{-1}(\mathcal{E}_{j'}^s, p)$, in this example (2). For more details see [30].

https://doi.org/10.1371/journal.pone.0218776.g003

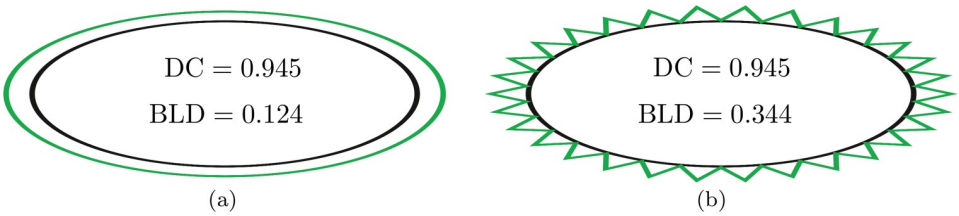

Fig 4. The robustness of the BLD in comparison to the DC for contour recognition. Here, the value of the surface delimited by the green contour is the same in both cases: (a) The region which should be recognized is an ellipse (black). (b) While the original contour was not recognized well at all, the DC for such a segmentation has yet the same high value as in (a). Using the BLD, we achieve a fairer evaluation of the segmentation, as the bad contour detection is taken into account and penalized with a higher value of the BLD (which corresponds to a minor similarity).

https://doi.org/10.1371/journal.pone.0218776.g004

Let S be the number of B-scans to be segmented by each expert and let $\mathcal{E}_j = \{\mathcal{E}_j^s\}_{s=1}^S$ and $\mathcal{E}_{j'} = \{\mathcal{E}_{j'}^s\}_{s=1}^S$ denote the segmentations done by the experts j and j' respectively. We define

$$\mathrm{BLD}(\mathcal{E}_j, \mathcal{E}_{j'}) = \sum_{s=1}^{S} \sum_{p \in \mathcal{E}_j^s} \frac{\mathrm{BLD}(p, \mathcal{E}_{j'}^s)}{n \max\limits_{p \in \mathcal{E}_j^s} \{\mathrm{BLD}(p, \mathcal{E}_{j'}^s)\}}, \tag{3}$$

where

$$\mathrm{BLD}(p, \mathcal{E}_{j'}^s) = \max\{d_{\min}(p, \mathcal{E}_{j'}^s), d_{\max}^{-1}(\mathcal{E}_{j'}^s, p)\},$$

and $d_{\min}(p, \mathcal{E}_{j'}^s)$ corresponds to the minimum distance from a point p on the reference $\mathcal{E}_j$ to the test contour $\mathcal{E}_{j'}^s$, while

$$d_{\max}^{-1}(\mathcal{E}_{j'}^s, p) = \max_{q \in \mathcal{E}_{j'}^s} \{d_{\min}(q, \mathcal{E}_j^s) \mid d_{\min}^{-1}(q, \mathcal{E}_j^s) = ||q - p||_2\}$$

denotes the maximum inverse distance at p on $\mathcal{E}_j^s$, as illustrated in Fig 3.

Although BLD is more robust than Dice and Jaccard for the detection of contours, (see Fig 4) it is, like Dice and Jaccard, not suited for the algorithm we present. As mentioned above, this algorithm is registration- but not segmentation-based, and its result is the displacement field over time of the entire border area between sclera and choroid (including the exact CSI, which is very difficult to localize). Therefore, the algorithm does not provide contours and thus, BLD, Jaccard, and Dice are not suitable as metric because we compare neither overlapping surfaces nor contours.

Consequently, we need to introduce an appropriate metric for our task, as of now diffZ (see Eq (4)) consisting of the difference between the detected displacements of the algorithm and the differences between the expert segmentations at different times, i.e.

$$\mathrm{diffZ}(\Delta\mathcal{E}_j, \Delta\mathcal{E}_{j'}) = \frac{\sum_{s=1}^{S} \sum_{i=1}^{n} |\Delta\mathcal{E}_j^s(i) - \Delta\mathcal{E}_{j'}^s(i)|}{mnS}, \tag{4}$$

where $\Delta\mathcal{E}_j = \{\Delta\mathcal{E}_j^s\}_{s=1}^S$ and $\Delta\mathcal{E}_{j'} = \{\Delta\mathcal{E}_{j'}^s\}_{s=1}^S$ represent the displacements in anterior-posterior/z-direction detected by the expert j and j' respectively, and are computed as the difference between the segmentations of the second measurement and those of the first one. The metric diffZ provides a value between 0 and 1, denoting the normalized difference (or, in other words, the amount of disagreement) between the detected displacements of algorithm and one expert, or, between any two experts, respectively. As a result, 0 means "no difference" (or, maximal agreement at each point) and 1 "maximum difference" (or, complete disagreement),

respectively. The metric diffZ is especially suitable for the evaluation of registration-based algorithms such as CRAR, in which no contours are shown. At the same time, it can also be applied in longitudinal studies to evaluate segmentation-based algorithms, which aim at the segmentation of the CSI. In this case, diffZ can be defined for the algorithm as the difference between the segmentations of two measurements performed within a time interval.

Power analysis (paired t-test)

As an additional component to the presented validation framework, we now perform an extended power analysis. Let $\Delta\mathcal{E} = \{\Delta\mathcal{E}_j^s\}_{j=1,s=1}^{r,S}$ denote the total of all displacements detected by all experts. For each $j \in \{1, \ldots, r\}$ the artificial ground truth

$$\bar{\mathcal{G}}_j = \frac{1}{r-1}\sum_{j' \neq j}^{r} \Delta\mathcal{E}_{j'}, \quad \text{for all } \Delta\mathcal{E}_{j'} \in \Delta\mathcal{E} \setminus \{\Delta\mathcal{E}_j\}, \tag{5}$$

is defined by leaving out the results of expert j and calculating the average of the displacements detected by the remaining experts.

Let X denote the displacements detected by the algorithm. For each expert j a paired t-test is done to compare the errors $X_j = X - \bar{\mathcal{G}}_j$ and $Y_j = \Delta\mathcal{E}_j - \bar{\mathcal{G}}_j$. In other words, after defining an artificial ground truth, we compare the difference in the errors of both algorithm and experts in their detection of choroidal thickness changes. Thus, we test the null hypothesis that the pairwise differences $X_j - Y_j$ come from a normal distribution with mean equal to 0 at the $\alpha = 0.01$ significant level. In order to reject the null hypothesis, the result of the p-value must be smaller than α. While a p-value shows whether an effect exists, it will not reveal the size of the effect (it might be, that a smaller p-value has occurred only on the basis of a large sample size [31]). This is why, we report both statistical (the p-value) and substantive significance (effect size). Using the Cohen's distance d between both datasets, the effect size can be determined by calculating the mean difference between the two groups X_j and Y_j, and then dividing the result by the pooled standard deviation $\sqrt{\mathcal{S}}$, i.e.

$$\text{Cohen's } d_j = \frac{\mu(X_j) - \mu(Y_j)}{\sqrt{\mathcal{S}}}, \quad \text{where } \mathcal{S} = \frac{(|X_j| - 1)\cdot\sigma^2(X_j) + (|Y_j| - 1)\cdot\sigma^2(Y_j)}{|X_j| + |Y_j| - 2}, \tag{6}$$

where $|X_j|$ and $|Y_j|$ denote the sample sizes of X_j and Y_j respectively, while $\sigma(X_j)$ and $\sigma(Y_j)$ are their standard deviations. The necessity of this "leave-one-out" power analysis lies in the fact that, in our case, an artificially generated ground truth can only be represented in the form of a matrix, as to every OCT B-scan pair a corresponding ground truth for comparison has to be generated. This results for $\bar{\mathcal{G}}_j$ in a matrix of size $n \times S$ corresponding to the ground truth for the entire dataset. As a result, the values of the standard deviations for $X - \bar{\mathcal{G}}_j$ vs. $\Delta\mathcal{E}_j - \bar{\mathcal{G}}_j$ are different from those for X vs. $\Delta\mathcal{E}_j$. Consequently, the values of the effect size, quantified by Cohen's d, change as well.

Material

Subjects

Ninety OCT 3D-volume stack pairs of Chinese subjects, aged 8-18 and stemming from urban regions with a high prevalence of myopia, have been analyzed. Healthy subjects with good distant and near vision (monocular corrected visual acuity was equal or better than LogMAR 0.00), but no systemic and ocular diseases, ocular trauma or surgery, were recruited. For the

subjects aged 8-13, the spherical refractive errors were -1.00D to -5.00D and cylindrical power was not more than -1.50D. For the others, the spherical refractive errors were +0.75D to -3.00D and the cylindrical power was not more than -1.00D. Written consent was obtained from both volunteers and, when necessary, their parents. The study protocol was approved by the Human Subjects Ethics Sub-committee of The Hong Kong Polytechnic University and was conducted in adherence to the tenets of the Declaration of Helsinki.

Data acquisition

The volunteers were measured twice at different times: half of them their left eye, the other half their right one. This resulted in 180 OCT volume stacks, each consisting of 25 B-scans of 500 × 768 pixels. The pixel spacing in nasal-temporal/x- and superior-inferior/y-direction were set to 11.46 μm/pixel and to 245 μm/pixel respectively, the one in anterior-posterior/z-direction was set to 3.87 μm/pixel. The volunteers were divided into three groups: the first group was measured a second time after 3 months, the second after 8 months, the third after 14 months.

The images were acquired by an eye-tracking dual-wavelength OCT system operating simultaneously at the 870 and 1075 nm bands. This system was developed at the HuCE-opto-Lab at the Bern University of Applied Sciences in Biel before it was transferred and setup at the Hong Kong Polytechnic University's School of Optometry. For technical details on the OCT system used we refer to the S2 Appendix.

Manual expert segmentation

Six ophthalmologists were recruited as experts. The experts received access to a Java-based online tool with which they could draw the CSI line, using either the mouse (PC) or a pen (tablet), see Fig 2. The experts could chose to segment the CSI without processing or to activate a pre-processing consisting of a histogram equalization and an average filtering. In this evaluation step the focus lay on both the intra-rater reliability of each individual expert and the agreement among the experts.

The 90 volume stack pairs were randomly distributed among the six experts, in such a way that each expert got exactly 50 volume stack pairs and each pair was assigned to at least three different experts for CSI segmentation. This allowed to test the agreement between the experts. From each volume stack pair, eight B-scan pairs were chosen for manual expert segmentation of the CSI, three in the lower (no. 1, 3 and 6), three in the middle (no. 11, 13 and 16) and two in the upper region (no. 21 and 23). To test the intra-rater reliability, each expert unknowingly received three times the same scan pair. The number of lines to be segmented by each expert was 2400 (8 scan positions per volume stack × 2 measurements per volunteer × 50 volume stacks × 3 repetitions). Calculating a time of approx. 9 sec per line, this adds up to 6 h per expert.

Manual segmentation (consensus)

Manual segmentation was done according to the following mutual consensus: the lower border of the choroid was drawn without taking into account the interconnecting tissues that appear as humps on top of the slowly varying baseline. Also the shadow artifacts that are cast by the retinal and choroidal vasculature were ignored. Vessels inside the sclera were also disregarded. It was agreed that once the CSI came to an end near the optical nerve, the segmentation line would be continued in the same direction and with the same slope, as depicted in Fig 2.

Results and discussion

Fig 5 shows the intra-rater reliability (repeatability) of the results of each expert by having the same image segmented thrice. The six experts together show an averaged standard deviation of ±24.73 μm in rating the same image. It is remarkable, that the experts are often inconclusive about the position of the CSI in the temporal para- and perifoveal region opposite to the ocular nerve, see later Fig 6(c) and 6(d). This is due to the low contrast in the image acquisition in this area of the choroid which makes a clear identification of the CSI difficult. The mean IRC value achieved by the experts was 0.807, which is to be considered as a good result. It has to be pointed out that the lowest scores were reached by those experts who probably had experienced difficulties in handling the tool. The first version of our online tool has been continuously optimized based on the feedback of the raters. Despite the challenges, the IRC scores were almost as good as we wished for.

Fig 7 presents the values of the WI calculated between each individual expert and the other members of the group for all the images (using DC, JC, BLD and diffZ as metric). When we compare the segmentations performed by the experts, the results show an average WI of 0.9992 ± 0.0221. The weak scattering of the WI demonstrates that its value does not depend on the choice of the similarity measure. Fig 7 shows that there are no relevant differences between the values of the WI whether they are calculated on the base of diffZ or the other three metrics, so Dice, Jaccard and BLD are not indispensable to this kind of evaluation, and therefore no longer needed. This justifies the use of diffZ for our case even more. The comparison between

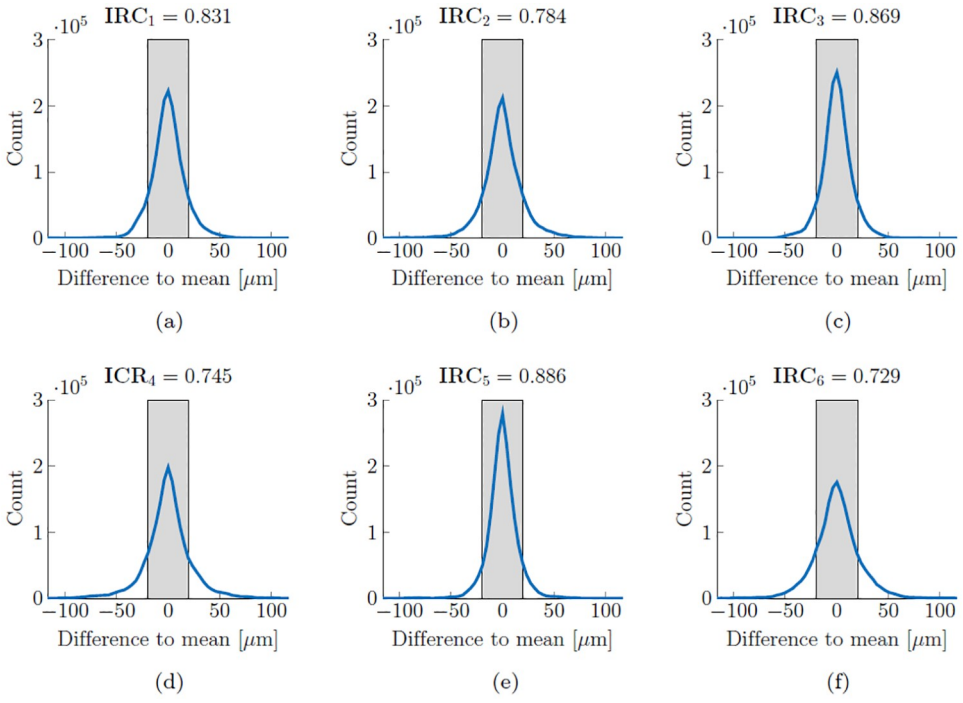

Fig 5. The representations of the intra-rater reliability of experts 1–6 ordered from (a)–(f). At every pixel position, the difference to the average value of the three available segmentations per rater is calculated. If its absolute value is smaller than a predefined threshold (here set to 20 μm represented by the grey area) then it is counted, i.e. the prediction is considered reliable. Therefore, the narrower and higher the curve, the more reliable the segmentation by the corresponding expert is. The number of counts found within this range is divided by the total number of segmentation points graded by the corresponding expert. By the obtained normalized value IRC$_j$ we define the Intra-Rater Coefficient to quantify the reliability of the j^{th} expert.

https://doi.org/10.1371/journal.pone.0218776.g005

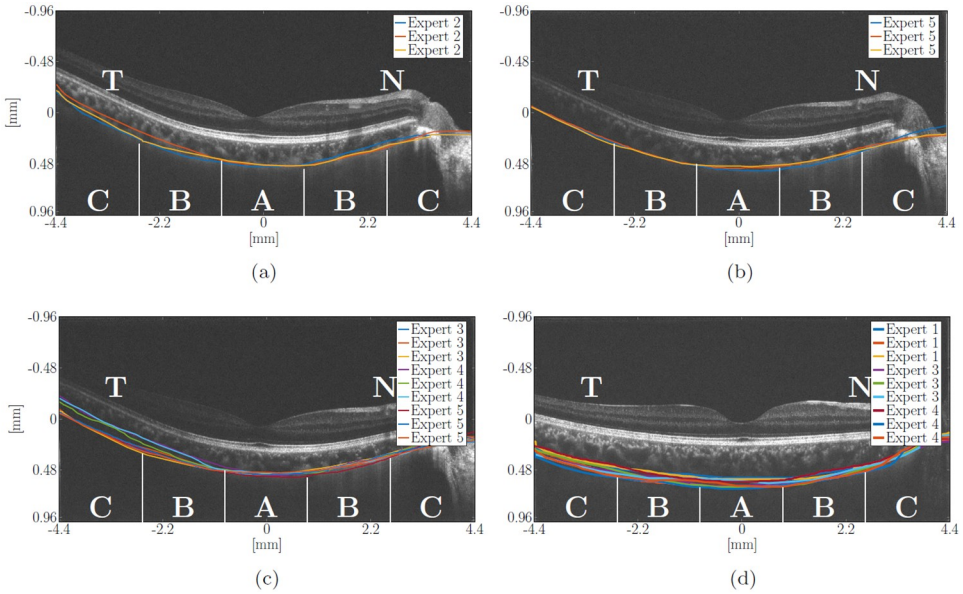

Fig 6. Examples of manual expert segmentation (consistent and less consistent with each other). Top: Repeatability of (a) expert 2 and (b) expert 5 when segmenting the CSI. Bottom: Comparison of segmentations by (c) experts 3, 4 and 5 and (d) experts 1, 3 and 4. The choroidal area is subdivided in nasal (N)-temporal (T)/*x*-direction into five equidistant regions (patches) symmetrically around the foveal center: A (foveal region), B (parafoveal region), and C (perifoveal region). Here only cases of right eyes are depicted.

https://doi.org/10.1371/journal.pone.0218776.g006

CRAR and the experts group shows that CRAR's predictions match those of the experts (the WI is greater than 1). To exclude that one expert influences the value of the WI much more than another, the WI is recalculated by omitting one expert at a time. The calculation of the WI between CRAR and the remaining five experts after leaving-one-out gives values in the range of [1.0126, 1.0393].

Fig 8 shows the variability at each scan position grouped by expert. As expected, the displacements detected by CRAR (4.29 ± 23.73 μm, see the far right hand side of Fig 8) are within the range of those detected by the experts (5.47 ± 39.45 μm) but with a smaller variance. This is

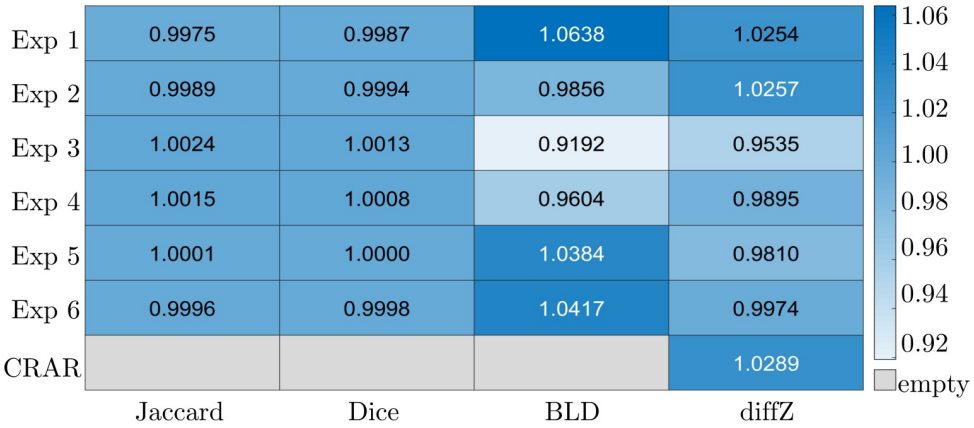

Fig 7. The values of the WI calculated for the algorithm and the experts group. As similarity measures Jaccard, Dice, BLD and diffZ are used.

https://doi.org/10.1371/journal.pone.0218776.g007

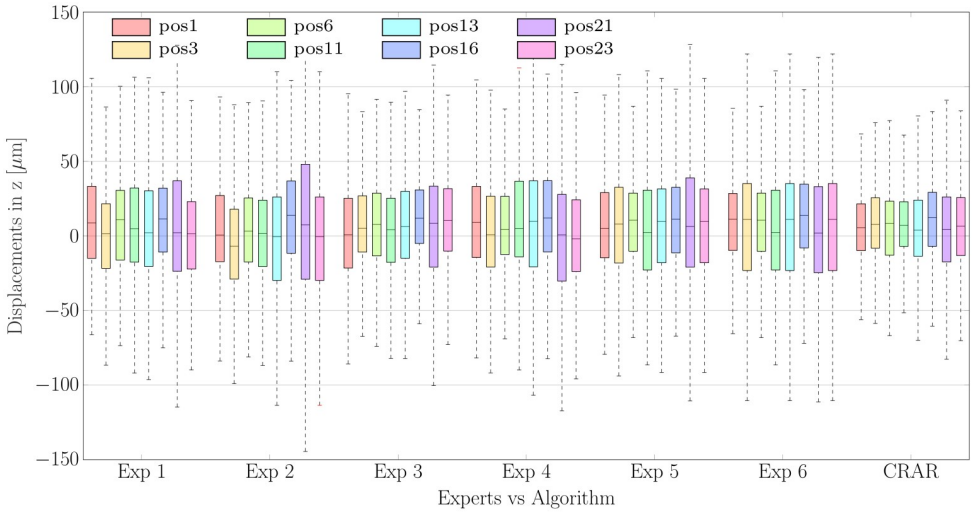

Fig 8. The average displacements of the CSI grouped by experts and algorithm. The results are obtained by manual segmentation by the six experts and by CRAR (subdivided into the B-scan positions 1, 3, 6, 11, 13, 16, 21 and 23).

https://doi.org/10.1371/journal.pone.0218776.g008

due to the fact that our automated algorithm guarantees perfect repeatability (IRC = 1), unlike the human rater.

Fig 9(a) and 9(b) show the total variability of all displacements detected by the experts and the algorithm, respectively, grouped by time interval between the two measurements. The average displacements measured using CRAR are 1.41 ± 16.23 μm in the case of images captured within a time interval of 3 months. For the other two time intervals of 8 and 14 months, the average changes are 4.88 ± 22.63 μm and 6.12 ± 29.32 μm, respectively. These results are still consistent with those of the experts, i.e. 1.76 ± 26.72 μm, 5.67 ± 32.48 μm and 7.62 ± 39.15 μm but with smaller variations. These results are also summarized in Table 1 for better data visualization. During each time interval between measurements the choroidal thickness increases, as is to be expected for growing tissue. This finding supports the hypothesis which several studies [7–10] formulate: the increase in thickness of the choroid is a normal feature of eye growth from early childhood to adolescence. Such a thickening of the choroid could relate to changes in the structure and associated physiological demands of the outer retina occurring during the eye's natural development, and/or to its role as sclera growth regulator [13].

On the other hand, the thickening of the choroid seems to slow down as time goes by, while the scattering of the data increases. The slower progress in thickening could be related to the high prevalence of Chinese children to become myopic (≈ 80% in the age 13–15). It could also be a sign that a disproportionate elongation of the eye-ball is taking place and must be compensated by slowing down the growth or, in case of myopia, even an actual thinning of the choroidal thickness at a relatively earlier age. The slowing down process of the thickening could also be the result of the natural "plateau" effect of the growth [10] in thickness of the choroid, in analogy to that of the body size. In other words, it appears that the thickness of the choroid increases in early childhood, reaching a peak between 10 and 20, and then exhibits a gradual decrease into older adulthood [8, 10]. However, it has to be noted that these assumptions are based on cross-sectional studies and, thus, the same subjects have not been observed regularly during longer periods of time. Therefore, before it can be generalized that the thickness of the choroid increases from early childhood to adolescence, further longitudinal research is ongoing, in which the subjects are being measured regularly and more frequently.

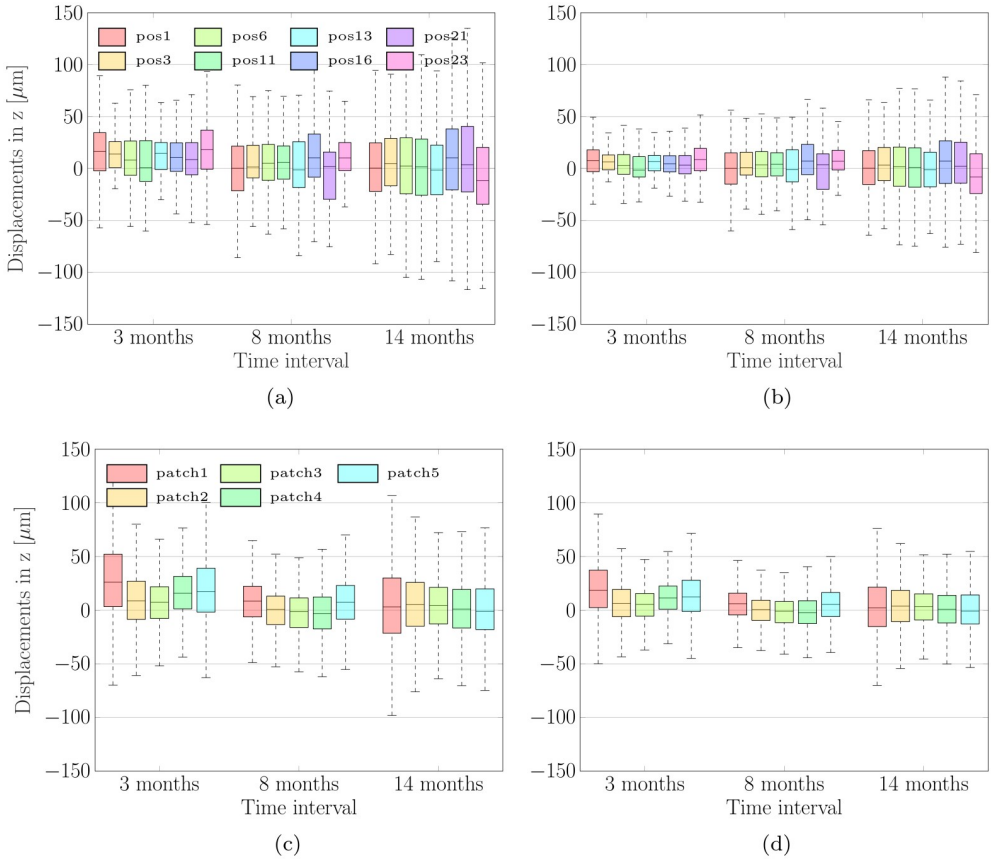

Fig 9. The average displacements of the CSI grouped by time intervals. Above: The average displacements of the CSI detected by the expert group (a) and by CRAR (b) grouped by time intervals between the two measurements and subdivided into eight scan positions. Below: The average displacements of the CSI detected by the expert group (c) and by CRAR (d) grouped by time interval and subdivided in nasal-temporal/*x*-direction into five equidistant regions C-B-A-B-C (patches) symmetrically distributed around the foveal center, see Fig 6.

https://doi.org/10.1371/journal.pone.0218776.g009

The time interval related increase in scattering of the detected displacements is a natural consequence of the diversity in choroidal changes, which, most likely, vary from subject to subject.

As mentioned above the results of CRAR are in line with those of other clinical studies: according to [9, 10] changes in the choroidal thickness appear to rapidly increase in early childhood (age 4–7, mean increase of 30 ± 15 μm within a time interval of 18 months), followed by a plateau in thickness change in the older age groups examined (age 10–13, mean change 13 ± 22 μm in 18 months). These aspects are reflected in our results and can be explained by the fact that the subjects that took part in this study are in an age phase, in which the choroid is still growing, but not as intensively as in early childhood (4–6 years old).

Fig 9(c) and 9(d) show the detected temporal changes grouped by time intervals and subdivided in nasal-temporal/*x*-direction into five equidistant regions C-B-A-B-C (patches) symmetrically distributed around the foveal center: A (foveal region), B (parafoveal region), and C (perifoveal region), see Fig 6. Here the mean changes detected by CRAR in the regions C-B-A-B-C (from the temporal (T) to the nasal (N) location in the case of a right eye, as shown in Fig 6, and in the opposite direction in the case of a left eye): 6.75 ± 27.12 μm (C), 4.07 ± 19.23 μm (B), 1.21 ± 13.57 μm (A), 4.23 ± 17.01 μm (B) and 5.18 ± 20.38 μm (C), respectively. These values are in the range of those detected by the experts: 9.95 ± 37.18 μm (C),

Table 1. The averaged vertical displacements Δz measured at different time intervals.

mean Δz in [μm] grouped by time interval			
	3 months	**8 months**	**14 months**
Experts	1.76 ± 26.72	5.67 ± 32.48	7.62 ± 39.15
CRAR	1.41 ± 16.23	4.88 ± 22.63	6.12 ± 29.32

The table notes the detected temporal changes depicted in Fig 9(a) and 9(b) averaged per time interval between the two measurements.

https://doi.org/10.1371/journal.pone.0218776.t001

Table 2. The vertical displacements Δz averaged per choroidal subregion.

Δz in [μm] averaged per choroidal subregion C-B-A-B-C					
	C(perifoveal)	**B(parafoveal)**	**A (foveal)**	**B (parafoveal)**	**C(perifoveal)**
Experts	9.95 ± 37.18	5.13 ± 28.53	1.81 ± 16.21	4.17 ± 22.37	6.27 ± 26.01
CRAR	6.75 ± 27.12	4.07 ± 19.23	1.21 ± 13.57	4.23 ± 17.01	5.18 ± 20.38

The table notes the detected temporal changes depicted in Fig 9(c) and 9(d) averaged per choroidal subregion C-B-A-B-C (i.e. the five equidistant patches symmetrically distributed around the foveal center).

https://doi.org/10.1371/journal.pone.0218776.t002

5.13 ± 28.53 μm (B), 1.81 ± 16.21 μm (A), 4.17 ± 22.37 μm (B) and 6.27 ± 26.01 μm (C), respectively. These results are also summarized in Table 2 for better data visualization.

According to [32, 33], the results show a prominent choroidal thickening in more peripheral regions and, in general, it can also be concluded that the first changes occur in the periphery rather than in the center. The error of CRAR is significantly lower than for the remaining five experts after leaving out one expert segmentation $\mathcal{E}_j$ at a time (p-value<0.01) with medium effect size (Cohen's d in the range of 0.41 and 0.49). This emphasizes the superior performance of CRAR in detecting choroidal thickness changes in comparison to those of the experts group.

The power analysis for different time intervals showed a small but not irrelevant effect size, and thus a change in the thickness of the choroid, which cannot be neglected. This supports the tendency mentioned above. For example, comparing the results for the time intervals of 3 and 14 months, we obtained a Cohen's d of 0.20 for the experts and 0.25 for the algorithm.

Conclusion

In this paper, we presented a statistical framework for validation of choroidal thickness changes detection without ground truth. Using the William's Index, we evaluated the agreement of experts' segmentations and illustrated, with CRAR as example, how to assess their performance in absence of a ground truth for comparison. With the help of the developed framework we demonstrated that our method CRAR provides results which are in the range of those of the experts, but with a lower variability. In addition, we confirmed this outcome using a modified state-of-the-art evaluation procedure: based on the results of a paired t-test, we could attest a higher precision of CRAR for detecting minute thickness changes of the choroid.

It can be concluded that CRAR provides a consistent, automated, expert-level performance in recognizing and monitoring subtle choroidal changes before a disease can actually manifest itself. Thus, we want to further develop CRAR so that it can be applied in the prevention and observation of several diseases and their respective treatments. In the ongoing research we plan to add children between 4–6 years to our test group and to verify the algorithm's performance for a time span of three years, in order to gather more information about the influence

of age, height, refractive error and axial length on the thickness of the choroid and to exhaustively discuss the results based on our new framework.

All things considered, the proposed validation framework is suitable for analyzing automatic detection algorithms for choroidal thickness changes, but might also be used for other applications, where a ground truth is not available.

Supporting information

S1 Appendix. Background information about CRAR.
(PDF)

S2 Appendix. Hardware: Hydra-Spectralis.
(PDF)

S1 Table. Choroidal thickness changes measurements.
(ZIP)

Acknowledgments

We would like to express our great appreciation to the HuCE-optoLab's staff that were involved in building the two commercial class prototype OCT systems, as well as their maintenance at Bern University of Applied Sciences, especially Markus Stoller, Michael Peyer, David Luggen, Patrick Arnold, Matthias Mooser and Christian Burri. Furthermore, we are grateful for the commitment of the staff at the Hong Kong Polytechnic University's School of Optometry, who performed the challenging long term study with children and teenagers.

We are grateful to the participating ophthalmologists Dr. Filippo Simona, Dr. Mali Okada, Dr. Daniel Barthelmes, Dr. Simon Rothenbühler and Dr. Emanuel Ramos de Carvalho for their assistance in the manual segmentations; their contributions to our research have been very valuable. We would like to thank the team at our commercial partner Heidelberg Engineering, Ralf Kessler and Tilman Otto for supporting the changes to the Spectralis hardware.

Author Contributions

Conceptualization: Tiziano Ronchetti, Christoph Jud, Hendrik P. N. Scholl, Boris Považay, Philippe C. Cattin.

Data curation: Tiziano Ronchetti, Christoph Meier, Rachel Ka Man Chun, Quan Liu, Chi-Ho To, Boris Považay.

Formal analysis: Tiziano Ronchetti, Christoph Jud, Peter M. Maloca, Selim Orgül, Alina T. Giger, Rachel Ka Man Chun, Chi-Ho To, Boris Považay, Philippe C. Cattin.

Funding acquisition: Christoph Meier, Boris Považay.

Investigation: Tiziano Ronchetti, Christoph Meier, Rachel Ka Man Chun, Quan Liu, Chi-Ho To, Boris Považay.

Methodology: Tiziano Ronchetti, Christoph Jud, Peter M. Maloca, Alina T. Giger, Hendrik P. N. Scholl, Chi-Ho To, Boris Považay, Philippe C. Cattin.

Project administration: Christoph Meier, Chi-Ho To, Boris Považay.

Resources: Christoph Meier, Boris Považay.

Software: Tiziano Ronchetti.

Supervision: Christoph Jud, Boris Považay, Philippe C. Cattin.

Validation: Tiziano Ronchetti, Christoph Jud, Peter M. Maloca, Philippe C. Cattin.

Visualization: Tiziano Ronchetti, Peter M. Maloca, Selim Orgül, Rachel Ka Man Chun, Quan Liu.

Writing – original draft: Tiziano Ronchetti, Christoph Jud, Boris Považay, Philippe C. Cattin.

Writing – review & editing: Tiziano Ronchetti, Christoph Jud, Peter M. Maloca, Selim Orgül, Alina T. Giger, Hendrik P. N. Scholl, Rachel Ka Man Chun, Boris Považay, Philippe C. Cattin.

References

1. Nickla DL, Wallman J. The multifunctional choroid. Progress in retinal and eye research. 2010; 29 (2):144–168. https://doi.org/10.1016/j.preteyeres.2009.12.002 PMID: 20044062
2. Chhablani J, Wong IY, Kozak I. Choroidal imaging: A review. Saudi Journal of Ophthalmology. 2014; 28 (2):123–128. https://doi.org/10.1016/j.sjopt.2014.03.004 PMID: 24843305
3. Chakraborty R, Read SA, Collins MJ. Diurnal variations in axial length, choroidal thickness, intraocular pressure, and ocular biometrics. Investigative ophthalmology & visual science. 2011; 52(8):5121–5129. https://doi.org/10.1167/iovs.11-7364
4. Ikuno Y, Kawaguchi K, Nouchi T, Yasuno Y. Choroidal thickness in healthy Japanese subjects. Investigative ophthalmology & visual science. 2010; 51(4):2173–2176. https://doi.org/10.1167/iovs.09-4383
5. Park KA, Oh SY. Choroidal thickness in healthy children. Retina. 2013; 33(9):1971–1976. https://doi.org/10.1097/IAE.0b013e3182923477 PMID: 23644561
6. Kim M, Kim SS, Koh HJ, Lee SC. Choroidal thickness, age, and refractive error in healthy Korean subjects. Optometry and Vision Science. 2014; 91(5):491–496. https://doi.org/10.1097/OPX.0000000000000229 PMID: 24727822
7. Mapelli C, Dell'Arti L, Barteselli G, Osnaghi S, Tabacchi E, Clerici M, et al. Choroidal volume variations during childhood. Investigative ophthalmology & visual science. 2013; 54(10):6841–6845. https://doi.org/10.1167/iovs.13-12761
8. Read SA, Collins MJ, Vincent SJ, Alonso-Caneiro D. Choroidal Thickness in Childhood. Investigative ophthalmology & visual science. 2013; 54(5):3586–3593. https://doi.org/10.1167/iovs.13-11732
9. Bidaut-Garnier M, Schwartz C, Puyraveau M, Montard M, Delbosc B, Saleh M. Choroidal thickness measurement in children using optical coherence tomography. Retina. 2014; 34(4):768–774. https://doi.org/10.1097/IAE.0b013e3182a487a4 PMID: 24013259
10. Read SA, Alonso-Caneiro D, Vincent SJ, Collins MJ. Longitudinal changes in choroidal thickness and eye growth in childhood. Investigative ophthalmology & visual science. 2015; 56(5):3103–3112. https://doi.org/10.1167/iovs.15-16446
11. He X, Jin P, Zou H, Li Q, Jin J, Lu L, et al. Choroidal thickness in healthy Chinese children aged 6 to 12: The Shanghai Children Eye Study. Retina. 2017; 37(2):368–375. https://doi.org/10.1097/IAE.0000000000001168 PMID: 27429378
12. Wang D, Chun RKM, Liu M, Lee RPK, Sun Y, Zhang T, et al. Optical defocus rapidly changes choroidal thickness in schoolchildren. PloS one. 2016; 11(8):e0161535. https://doi.org/10.1371/journal.pone.0161535 PMID: 27537606
13. Summers JA. The choroid as a sclera growth regulator. Experimental eye research. 2013; 114:120–127. https://doi.org/10.1016/j.exer.2013.03.008 PMID: 23528534
14. Vuong VS, Moisseiev E, Cunefare D, Farsiu S, Moshiri A, Yiu G. Repeatability of choroidal thickness measurements on enhanced depth imaging optical coherence tomography using different posterior boundaries. American journal of ophthalmology. 2016; 169:104–112. https://doi.org/10.1016/j.ajo.2016.06.023 PMID: 27345731
15. Ho M, Liu DT, Chan VC, Lam DS. Choroidal thickness measurement in myopic eyes by enhanced depth optical coherence tomography. Ophthalmology. 2013; 120(9):1909–1914. https://doi.org/10.1016/j.ophtha.2013.02.005 PMID: 23683921
16. Považay B, Hermann B, Unterhuber A, Hofer B, et al. Three-dimensional optical coherence tomography at 1050nm versus 800nm in retinal pathologies: enhanced performance and choroidal penetration in cataract patients. Journal of biomedical optics. 2007; 12(4):041211–041211. https://doi.org/10.1117/1.2773728 PMID: 17867800

17. Williams GW. Comparing the joint agreement of several raters with another rater. Biometrics. 1976; p. 619–627. https://doi.org/10.2307/2529750 PMID: 963175

18. Ronchetti T, Maloca P, Jud C, Meier C, Orgül S, Scholl HP, et al. Detecting Early Choroidal Changes Using Piecewise Rigid Image Registration and Eye-Shape Adherent Regularization. In: Fetal, Infant and Ophthalmic Medical Image Analysis. Springer; 2017. p. 92–100.

19. Kajić V, Esmaeelpour M, Považay B, Marshall D, Rosin PL, Drexler W. Automated choroidal segmentation of 1060 nm OCT in healthy and pathologic eyes using a statistical model. Biomedical optics express. 2012; 3(1):86–103. https://doi.org/10.1364/BOE.3.000086 PMID: 22254171

20. Tian J, Marziliano P, Baskaran M, Tun TA, Aung T. Automatic segmentation of the choroid in enhanced depth imaging optical coherence tomography images. Biomedical optics express. 2013; 4(3):397–411. https://doi.org/10.1364/BOE.4.000397 PMID: 23504041

21. Chiu SJ, Li XT, Nicholas P, Toth CA, Izatt JA, Farsiu S. Automatic segmentation of seven retinal layers in SDOCT images congruent with expert manual segmentation. Optics express. 2010; 18(18):19413–19428. https://doi.org/10.1364/OE.18.019413 PMID: 20940837

22. Beaton L, Mazzaferri J, Lalonde F, Hidalgo-Aguirre M, Descovich D, Lesk M, et al. Non-invasive measurement of choroidal volume change and ocular rigidity through automated segmentation of high-speed OCT imaging. Biomedical optics express. 2015; 6(5):1694–1706. https://doi.org/10.1364/BOE.6.001694 PMID: 26137373

23. Chen Q, Fan W, Niu S, Shi J, Shen H, Yuan S. Automated choroid segmentation based on gradual intensity distance in HD-OCT images. Optics express. 2015; 23(7):8974–8994. https://doi.org/10.1364/OE.23.008974 PMID: 25968734

24. Sui X, Zheng Y, Wei B, Bi H, Wu J, Pan X, et al. Choroid segmentation from optical coherence tomography with graph-edge weights learned from deep convolutional neural networks. Neurocomputing. 2017; 237:332–341. https://doi.org/10.1016/j.neucom.2017.01.023

25. Schlegl T, Seeböck P, Waldstein SM, Schmidt-Erfurth U, Langs G. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: International Conference on Information Processing in Medical Imaging. Springer; 2017. p. 146–157.

26. Grau V, Mewes A, Alcaniz M, Kikinis R, Warfield SK. Improved watershed transform for medical image segmentation using prior information. IEEE transactions on medical imaging. 2004; 23(4):447–458. https://doi.org/10.1109/TMI.2004.824224 PMID: 15084070

27. Rex DE, Shattuck DW, Woods RP, Narr KL, Luders E, Rehm K, et al. A meta-algorithm for brain extraction in MRI. NeuroImage. 2004; 23(2):625–637. https://doi.org/10.1016/j.neuroimage.2004.06.019 PMID: 15488412

28. Zijdenbos AP, Forghani R, Evans AC. Automatic "pipeline" analysis of 3-D MRI data for clinical trials: application to multiple sclerosis. IEEE transactions on medical imaging. 2002; 21(10):1280–1291. https://doi.org/10.1109/TMI.2002.806283 PMID: 12585710

29. Martin-Fernandez M, Bouix S, Ungar L, McCarley RW, Shenton ME. Two methods for validating brain tissue classifiers. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2005. p. 515–522.

30. Kim HS, Park SB, Lo SS, Monroe JI, Sohn JW. Bidirectional local distance measure for comparing segmentations. Medical physics. 2012; 39(11):6779–6790. https://doi.org/10.1118/1.4754802 PMID: 23127072

31. Sullivan GM, Feinn R. Using effect size—or why the P value is not enough. Journal of graduate medical education. 2012; 4(3):279–282. https://doi.org/10.4300/JGME-D-12-00156.1 PMID: 23997866

32. Park KA, Oh SY. Analysis of spectral-domain optical coherence tomography in preterm children: retinal layer thickness and choroidal thickness profiles. Investigative ophthalmology & visual science. 2012; 53 (11):7201–7207. https://doi.org/10.1167/iovs.12-10599

33. Ruiz-Moreno JM, Flores-Moreno I, Lugo F, Ruiz-Medrano J, Montero JA, Akiba M. Macular choroidal thickness in normal pediatric population measured by swept-source optical coherence tomography. Investigative ophthalmology & visual science. 2013; 54(1):353–359. https://doi.org/10.1167/iovs.12-10863