University of Dundee

PallorMetrics

Gibbon, Samuel; Muniz-Terrera, Graciela; Yii, Fabian S. L.; Hamid, Charlene; Cox, Simon; Maccormick, Ian J. C.

[Link to publication in Discovery Research Portal](#)

Download date: 04. Jul. 2024

tvst

PallorMetrics: Software for Automatically Quantifying Optic Disc Pallor in Fundus Photographs, and Associations With Peripapillary RNFL Thickness

Samuel Gibbon[1,2], Graciela Muniz-Terrera[3], Fabian S. L. Yii[1,2], Charlene Hamid[1], Simon Cox[5], Ian J. C. Maccormick[6,7], Andrew J. Tatham[1,8], Craig Ritchie[1,3], Emanuele Trucco[9], Baljean Dhillon[1,8], and Thomas J. MacGillivray[1,2,4]

[1] Centre for Clinical Brain Sciences, Edinburgh, UK
[2] Robert O Curle Ophthalmology Suite, Institute for Regeneration and Repair, University of Edinburgh, UK, Edinburgh, UK
[3] Centre for Dementia Prevention, University of Edinburgh, Edinburgh, UK
[4] VAMPIRE Project, Edinburgh Clinical Research facility, University of Edinburgh, Edinburgh, UK
[5] Lothian Birth Cohorts, Department of Psychology, University of Edinburgh, Edinburgh, UK
[6] Centre for Inflammation Research, University of Edinburgh, Edinburgh, UK
[7] Institute for Adaptive and Neural Computation, University of Edinburgh, Edinburgh, UK
[8] Princess Alexandra Eye Pavilion, Chalmers Street, Edinburgh, UK
[9] VAMPIRE Project, Computing (SSEN), University of Dundee, Dundee, UK

Correspondence: Samuel Gibbon, Chancellor's Building, The University of Edinburgh, 49 Little France Crescent, Edinburgh EH16 4SB, UK. e-mail: samuel.gibbon@ed.ac.uk

Purpose: We sough to develop an automatic method of quantifying optic disc pallor in fundus photographs and determine associations with peripapillary retinal nerve fiber layer (pRNFL) thickness.

Methods: We used deep learning to segment the optic disc, fovea, and vessels in fundus photographs, and measured pallor. We assessed the relationship between pallor and pRNFL thickness derived from optical coherence tomography scans in 118 participants. Separately, we used images diagnosed by clinical inspection as pale ($n = 45$) and assessed how measurements compared with healthy controls ($n = 46$). We also developed automatic rejection thresholds and tested the software for robustness to camera type, image format, and resolution.

Results: We developed software that automatically quantified disc pallor across several zones in fundus photographs. Pallor was associated with pRNFL thickness globally ($\beta = -9.81$; standard error [SE] = 3.16; $P < 0.05$), in the temporal inferior zone ($\beta = -29.78$; SE = 8.32; $P < 0.01$), with the nasal/temporal ratio ($\beta = 0.88$; SE = 0.34; $P < 0.05$), and in the whole disc ($\beta = -8.22$; SE = 2.92; $P < 0.05$). Furthermore, pallor was significantly higher in the patient group. Last, we demonstrate the analysis to be robust to camera type, image format, and resolution.

Conclusions: We developed software that automatically locates and quantifies disc pallor in fundus photographs and found associations between pallor measurements and pRNFL thickness.

Translational Relevance: We think our method will be useful for the identification, monitoring, and progression of diseases characterized by disc pallor and optic atrophy, including glaucoma, compression, and potentially in neurodegenerative disorders.

translational vision science & technology

Introduction

A pale optic disc is the hallmark of nonglaucomatous optic atrophy, which refers to the irreversible loss or damage of retinal ganglion cell axons along the anterior visual pathway.[1] A pale disc has numerous potential causes, including inflammation, ischemia, compression, increased intracranial pressure, toxicity, nutritional deficiency, trauma, hereditary conditions, vascular disease, infection, and retinal disease.[1,2] As such, disc pallor indicates the end stage of one of several disease processes. It begins to show approximately 4 to 6 weeks after axonal damage.[1] In clinical practice, a pale disc is often considered to be due to a compressive lesion until further tests prove otherwise.[1,3] Correctly identifying disc pallor can lead to life-saving treatment.

Pallor can be identified through ophthalmoscopy or fundus photography. However, these methods are limited in that assessing change over time may be difficult, judgment can vary substantially among observers,[4] and the location of pallor is often not recorded consistently.[5] Computational approaches may offer a solution, but efforts have been limited, either requiring special filters during acquisition[6–8] or manual demarcation of the disc.[9,10] The purpose of this study was to develop a fully automatic method of locating and quantifying disc pallor in fundus photographs.

The quantity of retinal ganglion cell axons can be observed directly with optical coherence tomography (OCT), which images the retinal nerve fiber layer (RNFL). Accordingly, we validated the tool by comparing pallor with peripapillary RNFL (pRNFL) thickness from OCT scans in anatomically equivalent zones. Additionally, we tested the software on an image set with clinically diagnosed pallor. Finally, we tested the robustness of the software to camera type, image format, and resolution with a variety of datasets.

Materials and Methods

Participants and Image Capture

Several datasets were used in model development and testing (Fig. 1); however, only one dataset (PREVENT) was used for testing the association between pallor and pRNFL thickness. The PREVENT-Dementia study protocol is described elsewhere.[11] Briefly, participants aged 40 to 59 years were recruited through multiple sources from five sites in the UK. Retinal imaging was conducted in a substudy at the Edinburgh site only (*n* = 123), which included fundus photography centered halfway between the optic disc and the macula (Fig. 2, right), with a nonmydriatic 45° field of view camera (CR-DGi; Canon USA, Inc., Lake Success, NY), and OCT (Heidelberg SPECTRALIS, Heidelberg Engineering, Germany) (Figs. 2 and 4 left most images). The N-site circular scan OCT module was used, set to high speed (1536 A-scans) with a target automatic real-time function of 100. Participants provided written informed consent, and the study was carried out in compliance with the Declaration of Helsinki. Six additional datasets were used in model development and testing (Table 1).
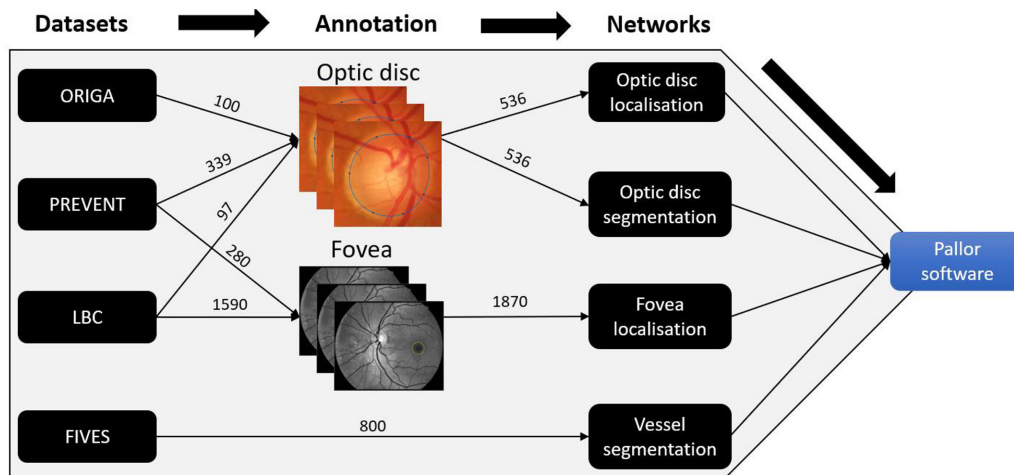

Figure 1. Flow of images through annotation, networks, and into the final software.
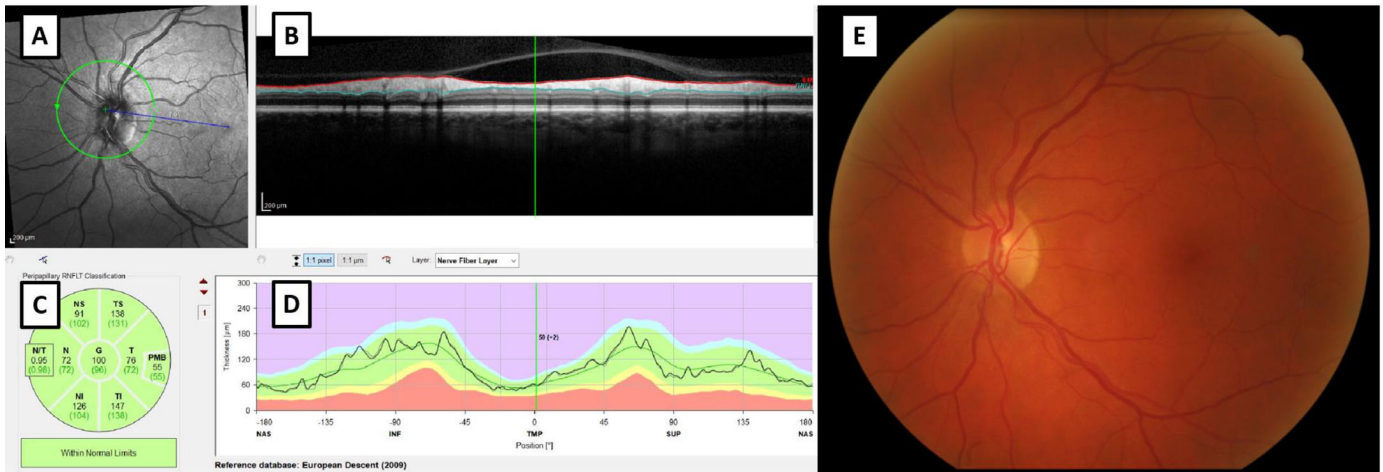
Figure 2. (**A**–**D**) Output of the Heidelberg SPECTRALIS peripapillary scan, where (**A**) shows the scan location, (**B**) shows the various layers (thick topmost layer is RNFL), (**C**) shows the measurement zones, and (**D**) shows the normative data (jagged line is the current participant). (**E**) shows the corresponding retinal fundus image.
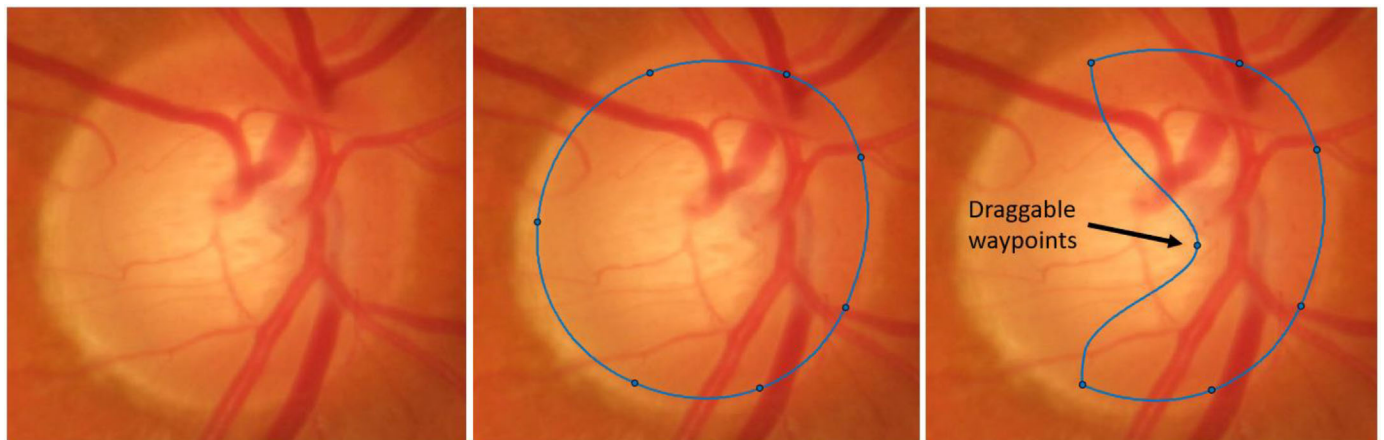

Figure 3. Annotation procedure. The annotator loads a full-sized image and zooms into the optic disc. The user then drags the waypoints of a deformable ellipse to the desired location. Additional waypoints can be added by double clicking. The entire shape can also be dragged. Performed in MATLAB with custom-written code.

Optic Disc Annotation

We annotated the optic disc in 536 images from three datasets (100 ORIGA, 339 PREVENT, 97 LBC). In fundoscopy, it is widely accepted that the disc margin lies at the inner edge of the border tissue, defined as collagenous tissue that arises from the sclera to join Bruch's membrane, forming a scleral cuff or lip between the optic nerve head and the choroid, which gives rise to its characteristic appearance as a yellow–white halo or crescent (Fig. 3).[19,20] Accordingly, the annotation protocol required dragging the waypoints of a deformable ellipse to the inner edge of the border tissue in full-resolution RGB fundus images (Fig. 3). Annotation was performed by a single researcher (author S.G., a PhD student in retinal image analysis), using custom MATLAB code.

To assess interannotator agreement, a second researcher (author F.Y., a PhD student in retinal image analysis and an optometrist) annotated a subset of 100 images (30 ORIGA, 40 PREVENT, 30 LBC) using the same protocol. The agreement metric was the mean intersection over union (mIoU), calculated as the area of overlap divided by the area of union. The mIoU has been shown to be a suitable measure of interannotator agreement for medical image segmentation tasks.[21] The mIoU between the two annotators was 0.942 (94.2%). Overall, F.Y. annotated a smaller area than S.G. (mean number of pixels for all images = 75,844 vs. 84,701); however, this difference was largely driven by a low
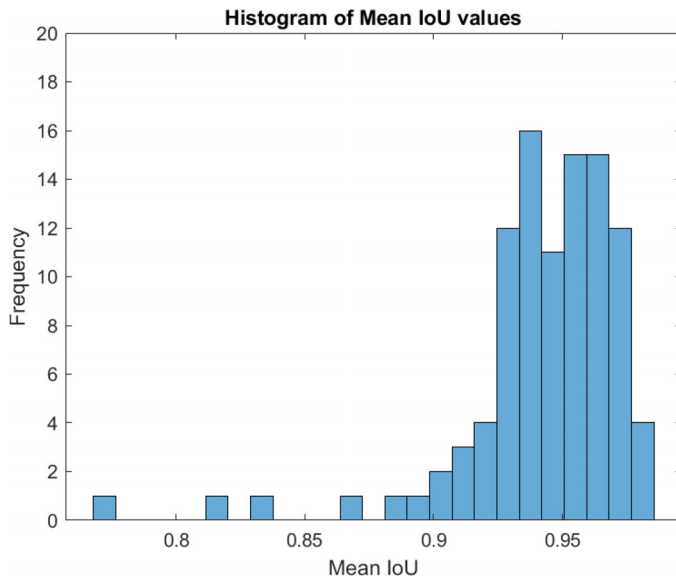
Figure 4. Histogram of interannotator agreement.

agreement on a few images (see the histogram in Fig. 4 and examples in Fig. 5).

Fovea Annotation

We annotated the fovea in 1870 images from the LBC and PREVENT datasets (280 PREVENT, 1590 LBC). The annotation procedure involved dragging a circle with a fixed radius of 150 pixels onto the estimated center of the fovea to generate a binary fovea map, where pixels inside the circle were labelled as fovea and pixels outside labelled as background. The image presented to the annotator was preprocessed to improve visibility by extracting the green channel of the RGB image and contrast was enhanced by performing contrast-limited adaptive histogram equalization. When the fovea was not visible (i.e., owing to poor illumination), the annotator estimated its position relative to the vessel arc of the central arcades and disc. If neither the vessel arc nor the disc was visible, the image was rejected. According to this protocol, 33 of 1623 images were rejected from the LBC, and none were rejected from PREVENT. Annotation was performed by a single researcher (author S.G.) using custom MATLAB code. No second annotator was used. The annotation and procedure are visualized in Figure 6.

Convolutional Neural Network Architecture and Computing Platform

Based on a 2022 survey of deep learning–based image segmentation,[22] we selected Google's DeepLabv3+ architecture,[23] which was the best performing network for image segmentation among the networks reviewed. DeepLabv3+ incorporates an encoding and decoding phase. The encoder–decoder model has been described elsewhere.[22,23] Briefly, in the encoding phase, information from the input image is extracted and compressed into a feature representation using a backbone convolutional neural network. The decoder then takes this as input to reconstruct the initial representation. The goal of such encoder–decoder architecture is for the model to learn a useful representation of the image. The result is accurate segmentation along object boundaries.[23] DeepLabv3+ can take one of several backbone architectures, including MobileNetv2[24] and Xception.[25] During experimentation, we found that MobileNetv2 produced the best results. Accordingly, we used DeepLabv3+ with a MobileNetv2 backbone pretrained on ImageNet[26] for all models, except vessel segmentation, for which we selected Xception. The values of relevant parameters varied with tasks and are given in the next section.

All models were trained in MATLAB (*version R2022b*; The MathWorks Inc., Natick, MA) using the Deep Learning Toolbox, on a Dell 7820 machine, fitted with an Intel Xeon Silver CPU and a NVIDIA Quadro RTX 5000 GPU, running Windows 10.

Optic Disc Localization

The optic disc localization network was trained on 536 images from three datasets (100 ORIGA, 339 PREVENT, 97 LBC). In preprocessing, we resized all images and their corresponding labels to 650×650 pixels, allowing the whole image to be processed by the network. We then split images into training, validation, and test sets, with a ratio of 80/10/10, yielding 429 images for training, 54 for validation, and 53 for testing. We used the Adam optimizer, the learning rate was constant at 0.0001, and the batch size was 4. During training, validation was carried out after every 100 iterations (approximately every epoch). We applied the following data augmentations to the training images to enhance generalization to unseen data: addition of random color jitter (brightness $= 0.3$, contrast $= 0.3$, saturation $= 0.3$), scaling (between a factor of 0.8 and 1.3), and rotation (between $-30°$ and $30°$). To prevent overfitting, we finalized training if the validation loss stopped decreasing or was equal to the previous 10 losses (validation patience).

The aim of the task was to locate the approximate center of the disc. There were far fewer pixels labelled as "disc" than pixels labelled as background (ratio $= 1/65$), leading to an imbalanced training set.

Table 1. Characteristics of Datasets Used in Model Development and Testing

Variable	PREVENT[12]	LBC 1936[13]	ORIGA[14]	FIVES[15]	RIM-ONE[16]	IDRiD[17]	RFMiD[18]
No. of images	339*	1623	650	800	169 (v1)485 (v3)	103	92
No. of images rejected	2	33	0	0	0	0	1
Camera	Canon CR-Dgi	Canon CR-Dgi	Canon CR-Dgi	Topcon TRC-NW8	Nidek AFC-210 & Kowa WX sD	Kowa VX-10α	TOPCON sD OCT-2000, Kowa VX-10α, TOPCON TRC-NW300
FOV	45	45	–	50	Nidek = 45, Kowa = 34	50	45–50
Resolution	3072 × 2048	2689 × 2186	2518 × 2048	2048 × 2048	Between 318 × 318 and 626 × 626	4288 × 2848	Between 2144 × 1424 and 4288 × 2848
Format	BMP	BMP	JPEG	PNG	PNG	JPG	PNG
Country	UK	UK	Singapore	China	Spain	India	India
Age, years (mean ± SD) or range	51.5 ± 5.5	72.5 ± 0.7	40–79	4–83	–	–	–
Female sex	57.70%	49.80%	–	–	–	–	–
Purpose of data collection	Cohort study in dementia	Population-based study of cognitive ageing	Population-based study of eye diseases	Enable AI-based methods of vessel segmentation	Glaucoma diagnosis	Diabetic retinopathy	To enable multidisease detection research

*This number includes images from professional rugby players, a substudy within Edinburgh, which were used in model development, but not analyzed further, partly because OCT was not captured in this group.

LBC 1936, Lothian Birth Cohort 1936; hereafter, LBC; ORIGA, Online Retinal Image Database for Glaucoma Analysis and Research; FIVES, Fundus Image VEssel Segmentation; RIM-ONE, Retinal IMage database for Optic Nerve Evaluation; IDRiD, Indian Diabetic Retinopathy Image Dataset; RFMiD, Retinal Fundus Multi-Disease Image Dataset); FOV, field of view); SD, standard deviation.
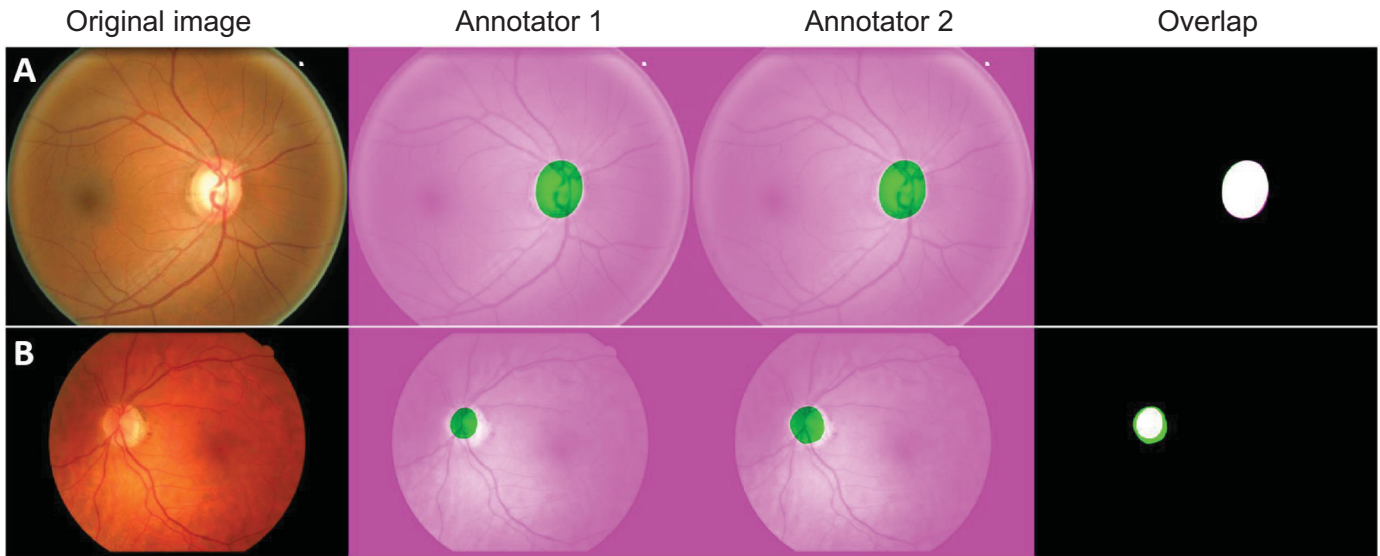
Original image	Annotator 1	Annotator 2	Overlap

Figure 5. Interannotator agreement. (**A**) A high level of agreement. (**B**) A low-level agreement. Annotator 1 (S.G.); annotator 2 (F.Y.).
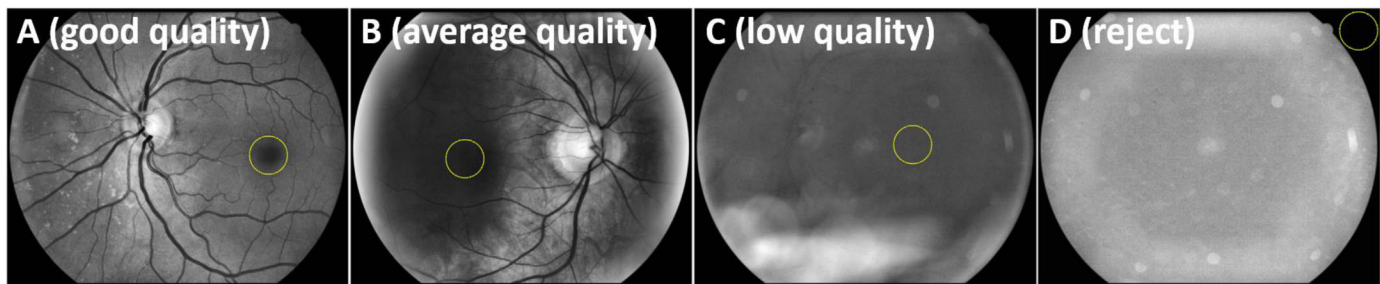

Figure 6. Fovea annotation procedure. (**A**) A good quality image, with the fovea clearly visible. (**B**) An area of low illumination over the macula, but the fovea can still be estimated. (**C**) Very low illumination and blur across the image; however. the optic disc and vessel arc are still visible, allowing the fovea to be estimated. (**D**) Neither the vessel arc nor the optic disc is visible; fovea estimation not possible.

With common metrics, a model would perform quite well if all pixels were simply labelled as background. To account for imbalance, we replaced the classification layer with a class weights layer, based on the class distribution of the full image set.

The convergence criterion was met after 3700 iterations (during the 28th epoch). The final validation accuracy was 99.16%. After postprocessing (removing the smallest object where multiple objects were detected), we computed the mean Euclidean distance (mED) in pixels between the ground truth and prediction based on the central points (Fig. 7). The mED in the test set was 2.06 ± 1.21 pixels, which expressed as a percentage of disc size (major axis length) was 2.02% $\pm$ 1.2%.

Optic Disc Segmentation

The optic disc segmentation network was trained on 536 images from three datasets (100 ORIGA, 339 PREVENT, 97 LBC). Input to the network was a $650 \times 650 \times 3$ RGB image, cropped around the disc, plus its corresponding ground truth segmentation. We split images into training, validation, and test sets, with a ratio of 80/10/10, respectively. We used the Adam optimizer, the learning rate was constant at 0.0001, and the batch size was 4. During training, validation was carried out after every 100 iterations (approximately every epoch). Augmentations were applied as before, but with additional levels of random color jitter (brightness = 0.6, contrast = 0.6, saturation = 0.6) and scaling (between a factor of 0.4 and 1.6). As before, class imbalance was high, albeit less than before (ratio $\approx$ 1/4.17), which we accounted for in the same way. The validation patience was 20. The convergence criterion was met after 16,400 iterations (during the 123rd epoch). The final validation accuracy was 98.79%. The mIoU for the test set (53 images) was 0.959 (95.9%). Two examples from the test set are presented in Figure 8.

Figure 7. Disc localization results on the test set. The ground truth is represented by a circle, and the prediction by an asterisk.
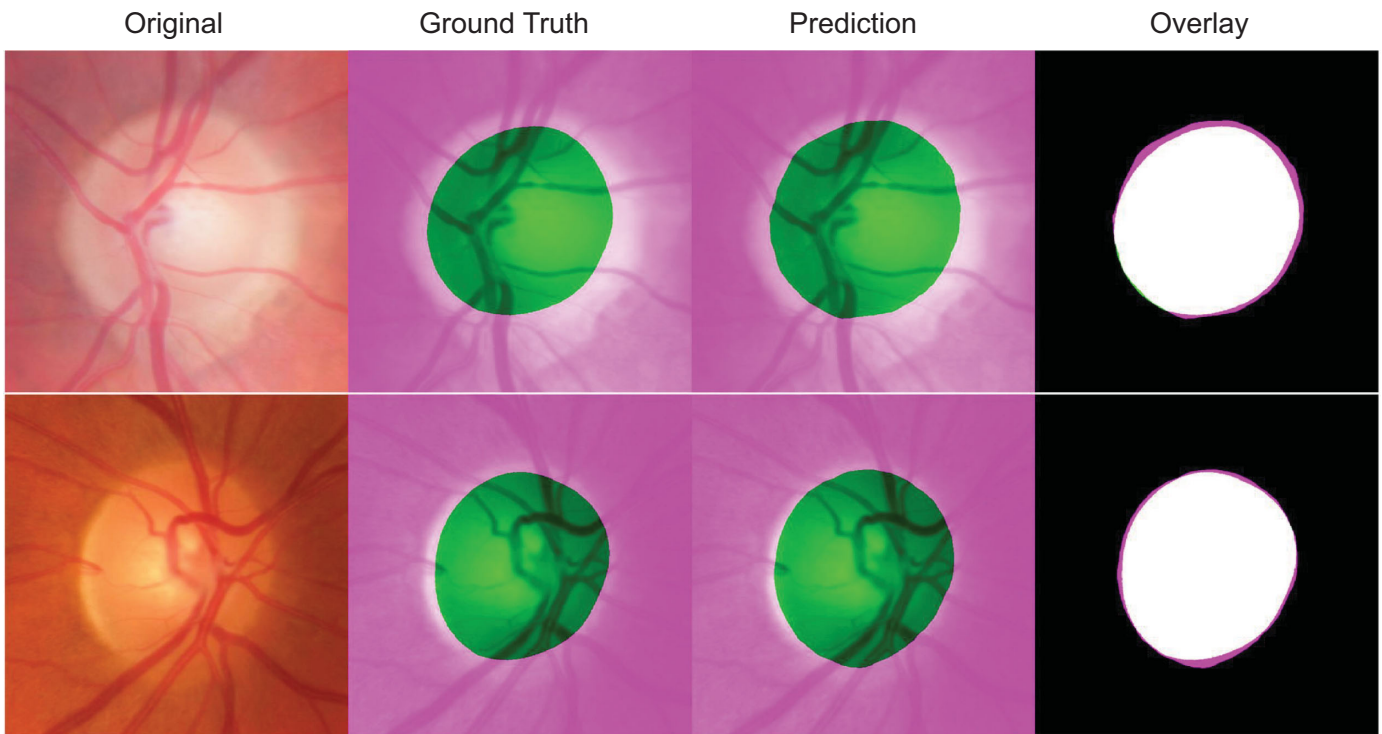
Figure 8. Optic disc boundary predictions from the test set.

We carried out external testing on IDRiD and RIM-ONE. Before generating predictions for IDRiD, each input image was resized, while maintaining its aspect ratio, to the median height of the training images. This strategy ensured that, when the image was cropped around the disc, the ratio of disc to background was approximately similar to that of a typical image from the training set. We carried out this step to improve generalizability.

The RIM-ONE dataset contains images that are already cropped around the disc; therefore, we skipped the disc localization stage. Our main performance metric was the mIoU; however, to enable a better comparison with other work, we also provide mean accuracy, defined as TP/(TP + FN), where TP is a true positive and FN is a false negative. We provide a direct comparison with the state of the art for IDRiD; however, for RIM-ONE we applied our network to their most recent data release (release 3). Other networks cited here have only been applied to release 1; therefore, comparative results are indicative only. On IDRiD, our model achieved a mIoU of 0.891, which beat the current state of the art (0.845). On RIM-ONE, we achieved a mIoU of 0.926 for nonglaucomatous

Table 2. Comparison of Our OD Segmentation Model to State of the Art

Method	Year	IDRiD (81 Images)		RIM-ONE V3 (Nonglaucoma; 313 Images)		RIM-ONE V3 (Glaucoma; 172 Images)		RIM-ONE V1 (169 Images)	
		mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU
ResNet + Unet	2020	–	–	–	–	–	–	–	0.925[a]
DRNet	2021	0.997	0.845	–	–	–	–	0.962	0.901
Ours	2022	0.899	0.891[b]	0.958	0.926[b]	0.947	0.907	–	–

ResNet + Unet,[28] DRNet.[29]
[a]Second best score.
[b]Best score.
mAcc, mean accuracy.

eyes, and 0.907 for glaucomatous eyes. A comparative summary is shown in Table 2.

Fovea Localization

The fovea localization network was trained on 1870 images from the LBC and PREVENT datasets (280 PREVENT, 1590 LBC). In preprocessing, we resized each image to 224 × 224 pixels. Unlike other features, the fovea is often not visible; however, its location can be inferred from its position relative to other salient features, including the vessel arc of the central arcades and the large dark patch covering the macula, which is common in poorly illuminated images. We hypothesized that substantially decreasing the image size would force the network to focus on these features. Therefore, input to the network was a 224 × 224 × 3 RGB image and its corresponding label. Images were split into training, validation, and test sets, with a ratio of 70/10/20, respectively, yielding 1309 images for training, 187 for validation, and 374 for testing. We chose a relatively large (20%) test set for this task as the focus was on generalization. We used the Adam optimizer, the learning rate was constant at 0.0001, and the batch size was 64. During training, validation was carried out after every 20 iterations.

We manually stopped training after 1167 iterations (during epoch 41), when we observed that the model was no longer improving. In postprocessing, we removed the smallest object, where multiple objects were detected. As with the optic disc localization network, we measured performance in the test set by calculating the Euclidian distance between the ground truth and prediction based on their central points. The mED was 3.77 ± 2.71. Expressed as a percentage of image height (224 pixels; disc size was not consistently available owing to image quality), the mED was 1.68% ± 1.11%. Several results from challenging images in the test set are presented in Figure 9.

We carried out external testing on IDRiD (103 images). In addition to providing the mED, we evaluate performance with the 1R criterion,[29] which refers to the radius of the optic disc. The 1R grid is centered on the fovea, and a score of 1 is given if the predicted coordinates lie within a given region (Fig. 10). The mED was 64.38 ± 76.43, and median Euclidean distance was 38.71. Overall, 95.15% of predictions fell within 1R, 86.41% within 0.5 R, 71.84% within 0.25 R, and 4.85% fell outside 1R (failed). The current state of the art for IDRiD is mED 41.87.[28]

Vessel Segmentation

We trained the vessel segmentation network on 800 images from the FIVES dataset. In preprocessing, we used the optic disc localization network, described elsewhere in this article, to crop each image and its counterpart vessel mask to 650 × 650 pixels centered on the disc. Therefore, input to the network was a 650 × 650 × 3 RGB image and its corresponding label. We split images into training, validation, and test sets, with a ratio of 70/15/15, resulting in 560 images for training, 120 for validation, and 120 for testing. Unlike our previous networks, we used Xception[25] as the backbone, because it generated more accurate segmentations during experimentation. We used the Adam optimizer; the learning rate was 0.0001, which we set to decrease by a factor of 0.1 in a piecewise manner every five epochs. The batch size was 4, and validation was carried out every 100 iterations. Augmentations applied were identical to those used for disc localization.

We manually stopped training after 3561 iterations (during epoch 6), because the model had converged. The final validation accuracy was 97.2%. The mean accuracy on the test set was 95.43% and the mIoU was 0.88. An example of the ground truth and automatic result is presented in Figure 11.
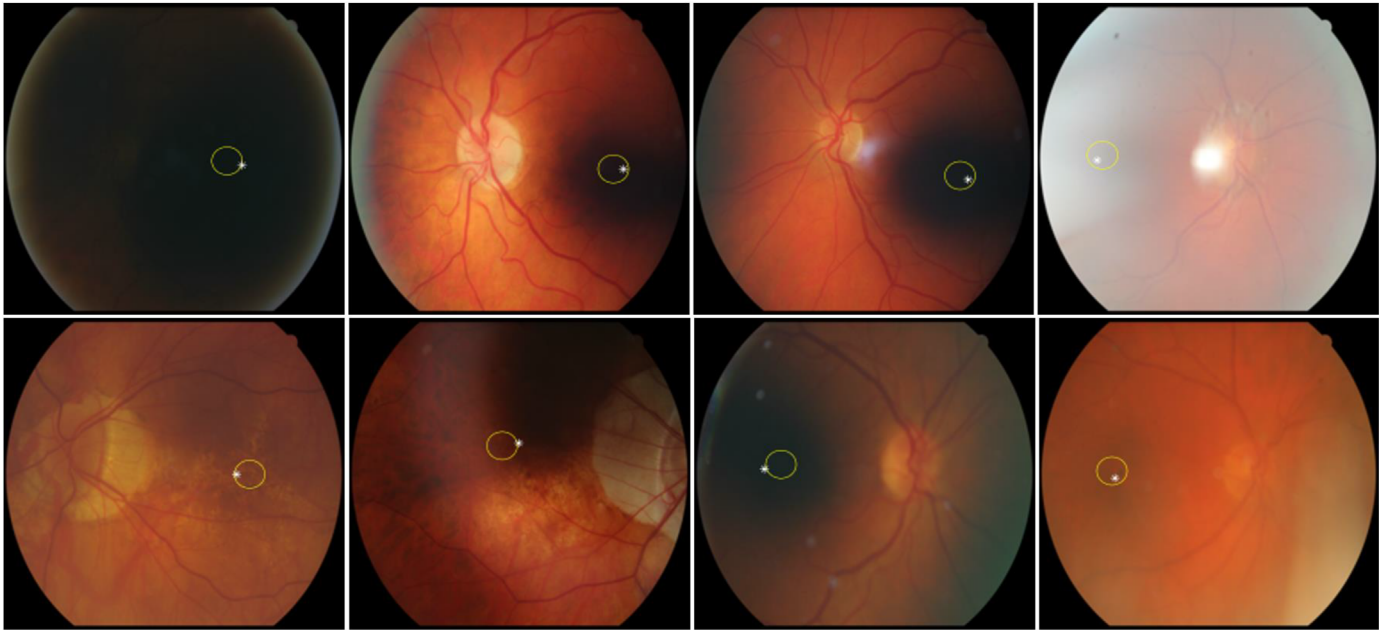
Figure 9. Fovea estimations in challenging images from the test set. The yellow circle represents the ground truth, and the asterisk is the prediction.
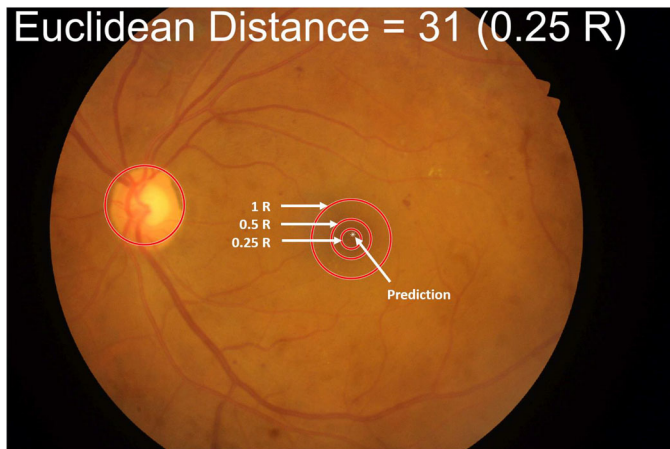

Figure 10. Both Euclidean distance and the 1R criterion were used to evaluate performance in the fovea detection network. The circular 1R grid is centered on the fovea. Image shown is from IDRiD.

Generating Pallor Measures

To calculate pallor, the software took as input full-size color fundus photographs. In preprocessing, a 300-pixel border of zeros was added to the left and right sides of the image, and the whole image was resized to the median height of the training images (2166 pixels), while maintaining its aspect ratio. Adding the border prevented a cropping failure when the disc was close to or on the border, and resizing helped the model to generalize by ensuring that the input image was approx-imately equivalent in size to the images the network was trained on.

The disc localization network was used to locate the disc center. Then, the image was cropped to a size of 650×650 pixels, with the disc at the center. Next, disc segmentation was performed on the cropped image. Postprocessing of the predicted disc boundary involved keeping the largest object (where multiple objects were detected), filling holes (where holes were detected), and smoothing edges. Edge smoothing consisted of three stages: (1) morphological opening with a disc-shaped structuring element of radius 75, (2) blur with a two-dimensional convolution, and (3) rethreshold to a value of 0.5. (Edge smoothing algorithm taken directly from MATLAB user "Image Analyst", available at https://uk.mathworks.com/matlabcentral/answers/380687-how-to-smooth-rough-edges-along-a-binary-image).

We defined the measurement region as starting at the inner edge of the border tissue and extending a fixed distance of 30 pixels inward. We chose this distance through direct observation as a balance between capturing as much of the neuroretinal rim (NRR) as possible while avoiding the cup. We defined the control region as starting at the outer border of the cropped image and extending a fixed distance of 50 pixels inwards. We chose this distance through direct observation as a compromise between capturing as much of the retina as possible while avoiding the disc and any atrophy. Vessels were detected in the cropped
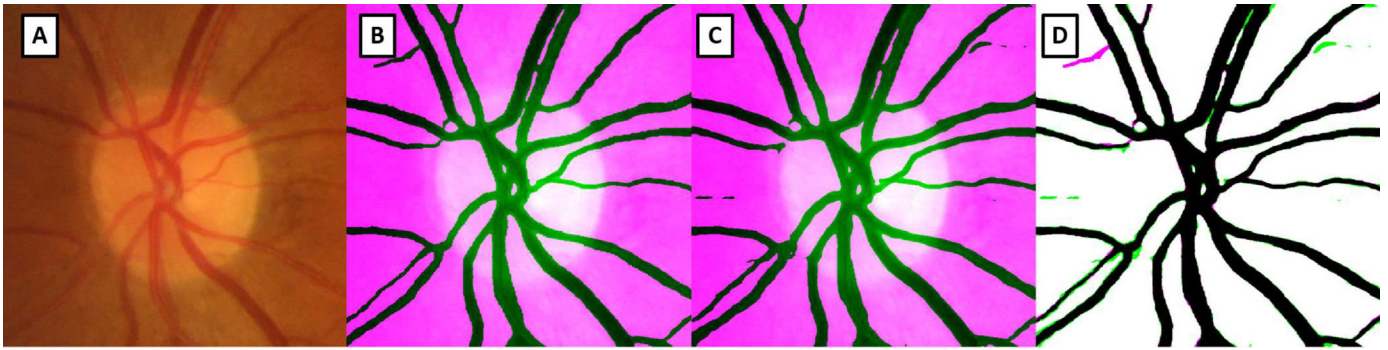
Figure 11. Vessel segmentation. (**A**) Input to the network, (**B**) ground truth, (**C**) automatic result, and (**D**) superposition of false-color image of ground truth and automatic result (false negative = green; false positive = magenta).

image and excluded (vessel pixels replaced with zero) from both the measurement region and the control region.

We then divided the measurement region into zones in accordance with the Heidelberg system for assessing pRNFL thickness. Specifically, the intersection of the optic disc–fovea axis and the measurement region took a value of zero degrees. The temporal zone then extended from 45° to −45°, the temporal inferior from 45° to 90°, and so on. The papillomacular bundle is a special case of the temporal zone, extending from 15° to −15°.

Finally, we calculated pallor based on the ratios of red and green pixel intensities[6,7,10,30,31] between the measurement and control regions. Specifically, we divided the mean of the green channel in the measurement zone by the mean of the red channel in the same zone. The result was then divided by the same measurement in the control region, except using the medians instead of the means. The result was a measure of pallor within each eye, for each zone.

Statistical Analysis

Data from one eye are correlated with data from the fellow eye.[32] Shuang et al.[33] showed that ignoring this intereye correlation in standard regression models can lead to spurious conclusions. The authors suggest that linear mixed effects modelling with the eye as the unit of analysis should be used. Accordingly, we modelled a random intercept for each person and eye. We adjusted *P* values for multiple comparisons with the false discovery rate procedure, which accounts for correlation between measurements. Statistical analysis was performed in R (version 4.2.1; www.R-project.org) using the lme4 and LmerTest packages.

To select covariates, we follow the dijunctive cause criterion,[34] which states that covariates should be added if they are causes of the exposure or outcome, or causes of both. Accordingly, the two retinal covariates are disc size and image brightness.

Disc Size

The NRR must accommodate between 0.9 and 1.5 million retinal ganglion cell axons. In a large disc, the axons can spread out, whereas in a small disc they are more compact. This means that, theoretically, the larger the disc, the paler it will appear, and vice versa. Indeed, we found a correlation between disc size and global pallor ($R^2 = 0.11$).

Brightness

We defined brightness as the median of all pixels inside the control region, after converting to greyscale, and removing vessels. Although the software controls for image brightness within each eye, light reflectance in the fundus is known to vary depending on the which part of the retina it strikes.[35] Specifically, the proportion of light reflected from the NRR and the background fundus may not remain constant with the level of light entering the pupil through the camera flash. Indeed, we found a correlation between brightness in the control region and global pallor ($R^2 = 0.12$).

Interocular Variability

We assessed interocular differences in pallor for each zone and propose a new measure, namely, interocular pallor variability, defined as the sum of absolute differences from all six zones.

Detecting Pallor in the RFMiD Dataset

We tested the results of our software on 92 images from the RFMiD dataset (images taken from the training set), one-half of which had been labelled as

optic disc pallor and one-half as disease risk = 0 by ophthalmologists from the RFMiD group. To assess group differences, we performed unpaired two-sampled Wilcoxon tests.

Testing the Software for Robustness to Camera System, Format, and Resolution

Among different countries, clinics, and research institutions, there is considerable heterogeneity in the technical aspect of retinal fundus images, including (i) the camera (e.g., Topcon, Canon), (ii) image resolution and image size, and (iii) file format (e.g., JPG, PNG, TIFF). To test the resilience of our system to these factors, a dataset was constructed containing images captured by various imaging systems, with different resolutions and formats. In addition, details pertaining to the field of view, centering protocol, and dilation were also noted. We chose 5 sets of 10 images from a total of 4 datasets (G1020[36], MESSIDOR[37] PREVENT, and REFUGE[38]). All these datasets, except for PREVENT, are accessible publicly. The task focus was on whether images could be processed successfully, not on how the software copes with images of varying levels of quality. Accordingly, we selected images with sufficient quality (broadly even illumination, free of major pathology). We judged the results by visual inspection, according to whether the software correctly (a) located the fovea, (b) located and segmented the disc, (c) rotated the image along the optic disc-fovea axis, and (d) segmented the vessels. We also recorded the computation time for each batch to assess whether processing time differed by dataset.

Developing a Set of Automatic Rejection Criteria

If there was insufficient information in an image to localize the disc or fovea, the image failed at the stage of processing. These images were usually very overexposed or underexposed (i.e., near totally white or black, respectively), or contained excessive blur. However, in most cases, the software processed the image, even when the quality was very low. To enable processing on large datasets, we aimed to develop a set of criteria through which such images could be rejected automatically. That is, although the images have been processed successfully, they are clearly not suitable for further analysis. For this task we used the LBC dataset, which contains images of varying levels of quality. We propose two automatic rejection criteria: disc eccentricity and control region brightness.

1. *Eccentricity* is the ratio of the distance between the center of an ellipse fitted onto the disc, and the major axis length, where 0 is a circle, and 1 is a line.
2. *Control region brightness* is the median of all pixels inside the control region, after converting to greyscale, and removing vessels.

By visual inspection, we aimed to develop conservative thresholds that would reject only the poorest quality images, or cases in which the software clearly failed for another reason (e.g., severe pathology).

Results

The software takes less than 3 seconds to process a single image and outputs several key visualizations (Fig. 12) alongside tabular data.

Quality Control and Sample Derivation

The sample derivation is illustrated in Figure 13. After quality control, concurrent fundus images and OCT scans were available for 118 participants (226 eyes). Three fundus images were rejected owing to segmentation error and low illumination (author S.G.; visual inspection) (Supplementary Fig. S1), and 13 OCT scans were excluded for reasons including clipping (4 images), improper centering (4 images), high myopia (≤ -5 diopters; 2 images), poor segmentation (3 images), poor illumination (1 image), and signs of pathology (5 images). Pathologies in OCT included epiretinal membrane, excessive peripapillary atrophy, and tilted discs. OCT quality control was carried out by C.H., an ophthalmic imager and analyst, via manual inspection of the images through the Heidelberg platform.

Summary Statistics

Image characteristics are summarized in Table 3, alongside basic demographics. Histograms showed that pallor was normally distributed (Supplementary Fig. S2).

Pallor was highest temporally and lowest nasally. Mean pallor was lower when considering all zones in the measurement region (global pallor; 1.37 ± 0.18) than when considering the entire disc (mean, 1.43 ± 0.2). In all measurement zones, pallor was numerically higher in the left eye compared with the right eye. pRNFL was thickest in both polar zones (superiorly and inferiorly), which was in accordance with typical findings.[39] Unlike pallor, pRNFL was not systemati-
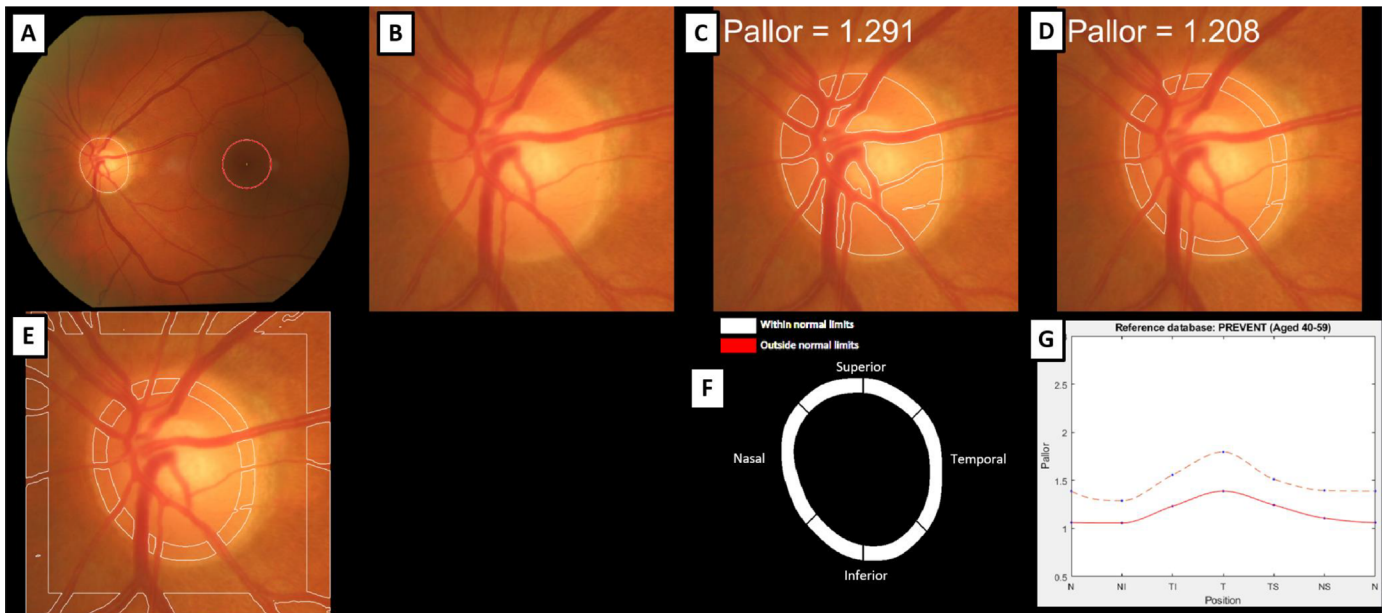
Figure 12. Core visualizations of the software. (**A**) Disc and fovea localization are used to rotate the image along the optic disc–fovea line. (**B**) Cropped optic disc. (**C**) Segmented disc excluding vessels. (**D**) Measurement region excluding vessels. (**E**) measurement and control region (outer square) excluding vessels. (**F**) Alert system (region lights up red if a limit is exceeded). (**G**) Dashed line represents 1 standard deviation above the mean of all participants in PREVENT, red line is the current participant.

cally different between the eyes. Boxplots of pallor and pRNFL by zone and eye are presented in Figure 14.

Associations Between pRNFL Thickness and Pallor

After adjusting for age, sex, disc area, control region brightness, and multiple comparisons, we observed statistically significant associations between pRNFL thickness and pallor globally ($\beta = -9.81$; SE $= 3.16$; $P < 0.05$), in the temporal inferior zone ($\beta = -29.78$; SE $= 3.32$; $P < 0.01$), and with the nasal/temporal ratio ($\beta = 0.88$; SE $= 0.34$; $P < 0.05$). Pallor in the measurement region was more discriminative than pallor measured in the whole disc ($\beta = -8.22$; SE $= 2.92$; $P < 0.05$). We also found an association between pRNFL thickness and pallor in the temporal-superior zone ($\beta = -17.29$; SE $= 7.83$; $P < 0.05$); however, this significance did not survive correction for multiple comparisons. Results are summarized in Table 4.

Interocular Pallor Variability

In the PREVENT dataset, data from both eyes were available for 108 participants. For global pallor, we measured a mean unit difference of 0.10 ± 0.07 between the eyes. To put this into context, global pallor ranges from 0.87 to 1.90. This result means that,

although interocular difference is evident, measurements from one eye in a person are broadly similar to measurements from the fellow eye. We also observed differences between zones, for example, the greatest difference between the left and right eyes was observed in the temporal region (mean, 0.13 ± 0.1). The general pattern of pallor being high temporally and low nasally was preserved between the eyes (Fig. 15). In Figure 16 (right), pallor is higher in the right eye than the fellow eye in the nasal zone, but higher in the left eye than the fellow eye in the temporal zone. To capture this zone-to-zone variability, we take the sum of absolute differences from all six zones. By contrast, in Figure 16 (left), pallor is higher in all zones in the left eye.

Assessing Pallor in the RFMiD Dataset

Of 92 images (46 pallor, 46 controls) in the RFMiD dataset (training set), which contained a patient group (diagnosed optic disc pallor) and healthy controls, the fovea localization module failed in one patient image and was subsequently rejected, despite accurate disc segmentation. Accordingly, analysis was carried out on 45 images labelled as pallor and 46 healthy controls.

Predicted pallor was substantially higher in the patient group compared with the control group for all zones. For example, the mean global pallor in the control group was 0.98 ± 0.09 compared with 1.23 ± 0.14 in the patient group, and this difference was statis-
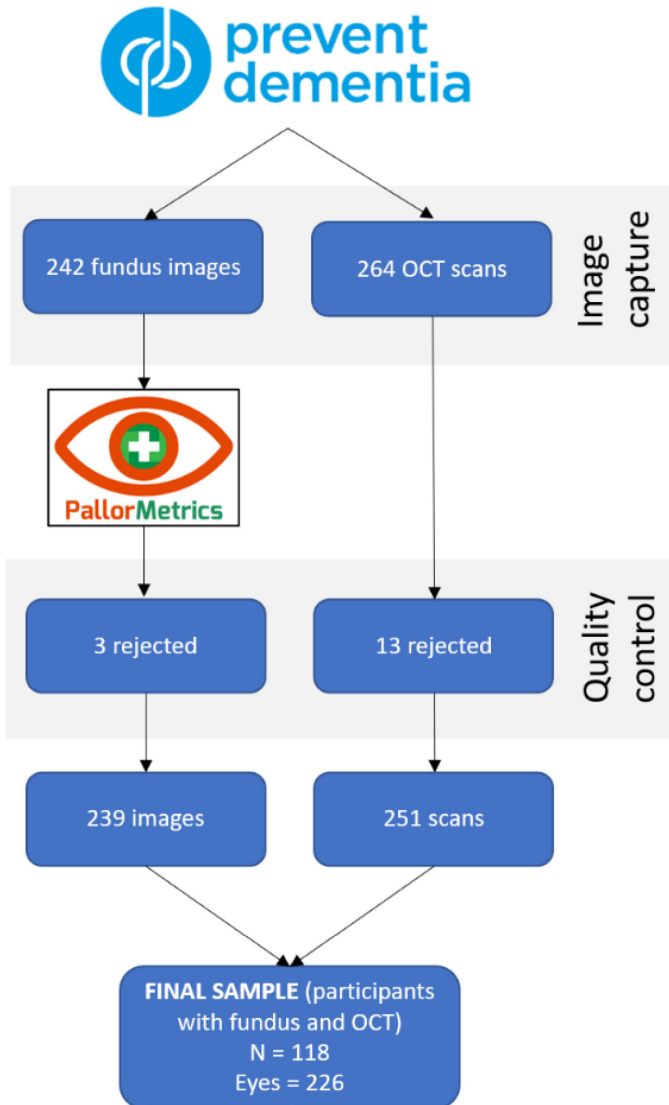
Figure 13. Sample derivation flowchart.

tically significant (Wilcoxon unpaired signed rank test: W = 208; $P < 10^{-11}$; R = 0.73) (Table 5). There was no evidence of significant difference in the nasal/temporal ratio between the groups (control group mean, 0.9 ± 0.07; patient group mean, 0.9 ± 0.05), reflecting the diffuse nature of pallor identified in the images. One example from each group is presented visually in Figure 17.

Robustness to Camera System, Format, and Resolution

We tested the software on five different datasets (none of which were used in model development) containing images captured with five different camera systems (from three manufacturers), three different image formats, and resolutions ranging from 1634 ×

1623 to 3072 × 2048. Judged by visual inspection, as per the criteria described in the Methods section, the methodology successfully processed all 50 images from all 5 datasets. Technical characteristics of the images are summarized in Table 6, and one example from each dataset is presented in Figure 18. Computation time was highest for the MESSIDOR dataset (2.9 seconds per image) and lowest for the REFUGE Canon dataset (2.5 seconds per image).

Developing a Set of Automatic Rejection Criteria

Of 1584 images from the LBC dataset, 13 failed processing for reasons including excessive blur, optic disc outside field of view, and overexposure and underexposure. Rejection thresholds were set based on visual inspection of the remaining 1571 images. Using our best judgment, thresholds for rejecting images were set at greater than 0.65 for eccentricity at greater than 0.65 and less than 50 for brightness of the control region. Examples of images that exceed these thresholds are presented in Figure 19. Summary statistics for images exceeding the proposed thresholds for both LBC and PREVENT are summarized in Table 7.

Discussion

We have presented a fully automatic method of quantifying optic disc pallor in color fundus photographs. In approximately 3 seconds per image, the software generates tabular data and visualizations capturing key measurements and summative properties. In particular, the software generates a global pallor metric, as well as metrics for seven zones, in accordance with the Spectralis OCT peripapillary scan. The software proved robust to camera system, image format, and resolution in our experiments and generates several metrics that can be used to filter out challenging or low-quality images, thereby allowing for application to large datasets.

In similar work, Yang et al.[30] developed a fully automatic pallor quantification system that operates on standard fundus photographs. However, their work has some limitations. For example, vasculature is included in their measurement region. This factor may be problematic, because vessel appearance is known to change with disease. For example, in hypertensive retinopathy, the arteriolar light reflex is accentuated,[40] in retinal vasculitis a white cuff is visible around vessels,[41] and, although rare, in lipemia retinalis, vessels appear creamy.[42] In addition, zones in Yang

Table 3. Demographics, Covariates, Pallor, and pRNFL Thickness in Microns by Zone and Eye

	Eye		
	Left	Right	Overall
N (participants)			118
N (eyes)	(n = 112)	(n = 114)	(N = 226)
Age, years	51.5 ± 5.60	51.4 ± 5.67	51.4 ± 5.63
Sex (female)	68 (60.7)	68 (59.6)	136 (60.2)
Disc area	75,300 ± 16,400	72,400 ± 14,500	73,900 ± 15,500
Luminance	93.1 ± 19.5	90.1 ± 18.8	91.6 ± 19.1
Temporal			
Pallor	1.62 ± 0.24	1.54 ± 0.22	1.58 ± 0.23
RNFL	69.3 ± 13.8	73.8 ± 13.1	71.6 ± 13.6
RNFL missing	5 (4.5)	6 (5.3)	11 (4.9)
Temporal inferior			
Pallor	1.42 ± 0.19	1.34 ± 0.19	1.38 ± 0.19
RNFL	139 ± 19.7	141 ± 22.7	140 ± 21.2
RNFL missing	5 (4.5)	6 (5.3)	11 (4.9)
Nasal inferior			
Pallor	1.16 ± 0.14	1.12 ± 0.14	1.14 ± 0.14
RNFL	113 ± 21.9	112 ± 22.1	113 ± 22.0
RNFL missing	5 (4.5)	6 (5.3)	11 (4.9)
Nasal			
Pallor	1.25 ± 0.16	1.19 ± 0.15	1.22 ± 0.18
RNFL	74.7 ± 16.2	76.9 ± 18.9	75.8 ± 17.6
RNFL missing	6 (5.4)	7 (6.1)	13 (5.8)
Nasal superior			
Pallor	1.27 ± 0.16	1.20 ± 0.15	1.23 ± 0.16
RNFL	111 ± 18.9	98.6 ± 18.8	105 ± 19.8
RNFL missing	6 (5.4)	7 (6.1)	13 (5.8)
Temporal superior			
Pallor	1.36 ± 0.19	1.30 ± 0.18	1.33 ± 0.19
RNFL	135 ± 17.3	135 ± 16.9	135 ± 17.0
RNFL missing	5 (4.5)	6 (5.3)	11 (4.9)
PMB			
Pallor	1.68 ± 0.25	1.59 ± 0.23	1.63 ± 0.25
RNFL	53.9 ± 13.1	55.6 ± 9.04	54.7 ± 11.2
RNFL missing	6 (5.4)	9 (7.9)	15 (6.6)
Global			
Pallor	1.41 ± 0.18	1.34 ± 0.17	1.37 ± 0.18
RNFL	98.1 ± 8.03	98.2 ± 8.12	98.2 ± 8.06
RNFL missing	6 (5.4)	7 (6.1)	13 (5.8)
Nasal/temporal ratio			
Pallor	0.77 ± 0.07	0.78 ± 0.08	0.78 ± 0.07
RNFL	1.13 ± 0.34	1.08 ± 0.36	1.10 ± 0.35
RNFL missing	7 (6.3)	10 (8.8)	17 (7.5)
Whole disc			
Pallor	1.46 ± 0.201	1.39 ± 0.185	1.43 ± 0.196

Values are mean ± SD or number (%).
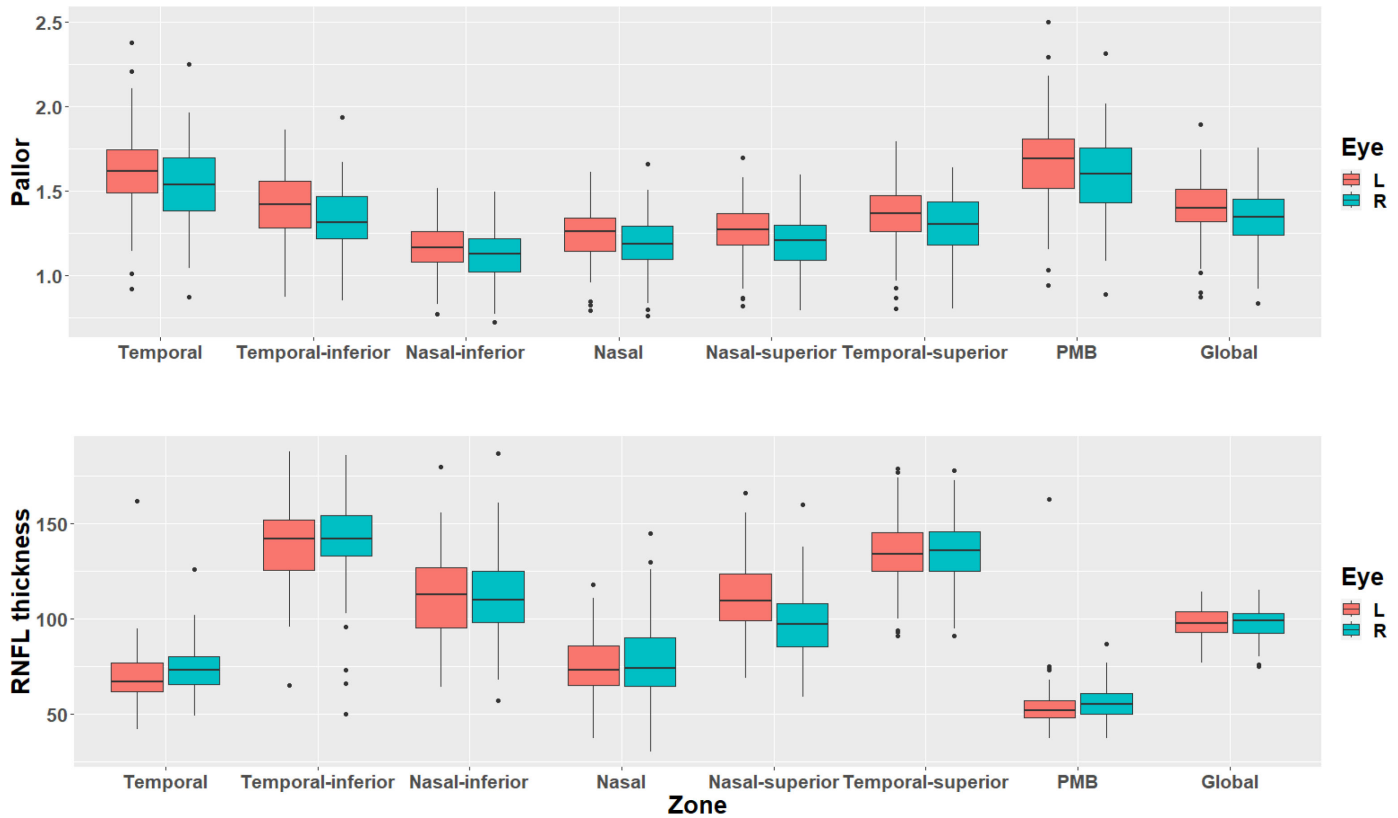PMB, papillomacular bundle.

Figure 14. Boxplots representing pallor and pRNFL thickness values by zone and eye. $N = 118$ (114 right eyes, 112 left eyes).

et al.'s work (clock-hour locations) were not defined by their spatial relation to the fovea, making it difficult to compare measurements accurately across different images or to make sectoral comparisons to OCT. Our approach addresses these limitations by (a) detecting vessels and excluding them from both the measurement and control region and (b) rotating the image along the optic disc–fovea axis before analysis.

In other similar work, Gonzalez-Hernandez et al.[31] developed a fully automated system to assess hemoglobin content in the optic disc (Laguna-ONhE; Optic Nerve Head Evaluation), which partly explains pallor. As with Yang et al.'s system, the Laguna software did not define the measurement zone in relation to the fovea. However, unlike Yang et al. and the current study, Laguna does attempt to segment the optic cup. Although this strategy carries the advantage of capturing the entire NRR (where possible), it may fail when the cup is not visible, which is often the case in fundus photographs. Indeed, numerous studies show that segmenting the cup is difficult,[4,27,43,44] although recent work has been more successful[45]; moreover, it is difficult to establish ground truth given interobserver variability in locating the extent of optic disc cupping. For this reason, we chose instead to

define the measurement region in accordance with Yang et al. at a fixed distance inward from the disc margin, sacrificing potential accuracy for robustness.

We investigated the relationship between pallor and pRNFL thickness in participants for whom concurrent data were available. Controlling for age, sex, disc area, control region brightness, and multiple comparisons, we found statistically significant associations between pallor and global pRNFL thickness, with a significant association also observed in the temporal-inferior and temporal-superior zones and in the temporal-nasal ratio. pRNFL thinning (as measured with OCT) is associated with several negative health outcomes, including glaucoma,[46] increased cardiovascular risk,[47] Alzheimer's disease, mild cognitive impairment,[48] future cognitive decline,[49] increased risk of dementia,[50] small vessel disease,[51] and stroke.[52] However, OCT is not yet widely available. Our approach generates measures of disc pallor that are associated with pRNFL thickness from simple color fundus photographs, which are much more widely available, potentially enabling the detection and monitoring of the progression of diseases that involve pRNFL loss with this imaging technology.

Table 4. Linear Mixed Effects Regression Models of pRNFL Thickness Predicted by Pallor in Equivalent Zones

Zone	Coefficients		
	β (SE)	P Value	P Value (Adjusted for FDR)
Global	−9.81 (3.16)	0.002**	0.011*
Temporal	−2.89 (4.68)	0.538	0.724
Temporal-inferior	−29.78 (8.32)	0.000***	0.004**
Nasal-inferior	5.92 (13.09)	0.652	0.724
Nasal	−4.7 (9.2)	0.610	0.724
Nasal-superior	−7.92 (9.85)	0.422	0.704
Temporal-superior	−17.29 (7.83)	0.028*	0.057
PMB	−0.03 (3.72)	0.994	0.994
Global pRNFL ∼ pallor in whole disc	−8.22 (2.92)	0.005**	0.018*
Nasal-temporal ratio	0.88 (0.34)	0.011*	0.028*

Random intercepts are modelled for each subject and eye. Models are adjusted for age, sex, disc area, and image brightness.
*$P < 0.05$.
**$P < 0.01$.
***$P < 0.001$.
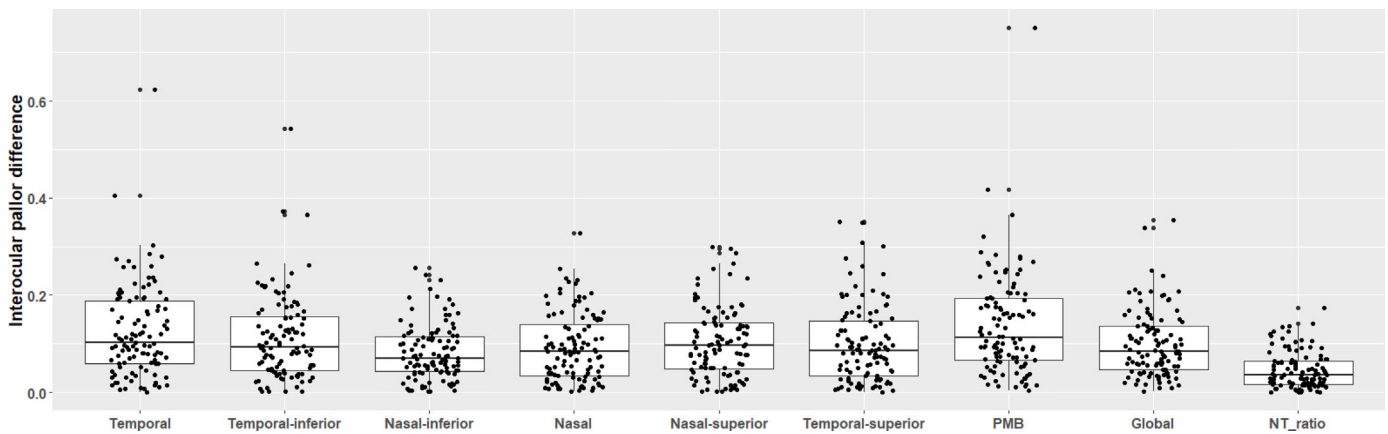FDR, false discovery rate; PMB, papillomacular bundle.

Figure 15. Boxplots representing the difference in pallor between the eyes of a participant ($N = 108$).

Aside from its association with pRNFL thickness, the ability to quantify pallor may have additional value in differentiating the etiology of structural changes to the optic nerve head; for example, in differentiating glaucomatous and nonglaucomatous optic neuropathy. Although pRNFL thinning is seen in both conditions, cupping rather than pallor is typical of glaucomatous optic neuropathy, and the presence of clinically apparent pallor often triggers investigations for nonglaucomatous causes, including potential magnetic resonance imaging of the anterior visual pathway.[6,53]

In all zones, pallor was slightly lower in the right eye compared with the left eye, and this finding largely corresponded with pRNFL thickness measured in equivalent zones. This observation is in agreement with other studies that found the RNFL to be consistently thicker in the right eye.[54–57] Cameron et al.[58] discussed the importance of interocular symmetry in health and disease, pointing out that the emergence of asymmetry may alert the ophthalmologist that glaucoma should be considered. Further, they review several studies that attempt to create thresholds for when RNFL asymmetry may be clinically meaningful for glaucoma diagnosis and progression. This observation further suggests that our measure of symmetry (interocular pallor variability), may find usefulness in glaucoma detection and diagnosis.

Another important use case for the software could be the identification, monitoring and progression of
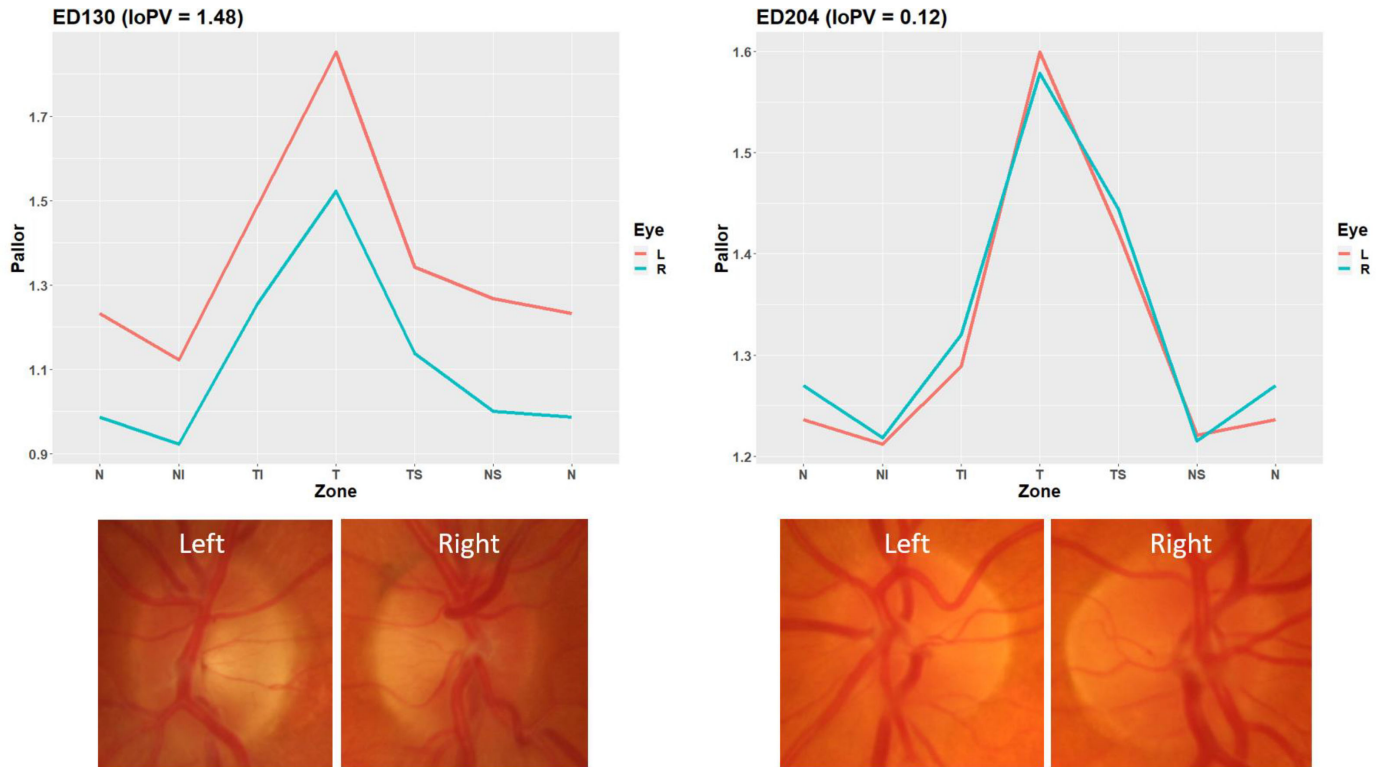
Figure 16. Parallel plots from two participants showing interocular differences in pallor by zone (nasal pallor is repeated on either side of each plot for aesthetics). I, inferior; IoPV, interocular pallor variability; N, nasal; S, superior; T, temporal.

Table 5. Unpaired Wilcoxon Signed Rank Test Results Comparing Eyes Labelled as Having Pallor vs Controls in the RMFiD Dataset

Zone	Group, Mean ± SD		Wilcoxon Test		
	Control	Pallor	W	*P Value*	R (Effect Size)
Temporal	1.06 ± 0.12	1.32 ± 0.15	195	$<10^{-10}$	0.7
Temporal-inferior	0.95 ± 0.09	1.2 ± 0.13	118	$<10^{-12}$	0.76
Nasal-inferior	0.91 ± 0.09	1.15 ± 0.13	143	$<10^{-11}$	0.74
Nasal	0.95 ± 0.1	1.18 ± 0.14	161	$<10^{-11}$	0.73
Nasal-superior	0.93 ± 0.09	1.14 ± 0.13	176	$<10^{-11}$	0.72
Temporal-superior	0.97 ± 0.1	1.24 ± 0.15	151	$<10^{-11}$	0.74
PMB	1.08 ± 0.12	1.34 ± 0.16	208	$<10^{-10}$	0.69
Global	0.98 ± 0.09	1.23 ± 0.14	164	$<10^{-11}$	0.73
Global pRNFL ~ whole disc pallor	1.03 ± 0.11	1.26 ± 0.15	217	$<10^{-10}$	0.68
Nasal-temporal ratio	0.9 ± 0.07	0.9 ± 0.05	1063	0.82	0.02

Note: We chose not to correct for multiple comparisons here owing to the very low *P* values.

PMB, papillomacular bundle; SD, standard deviation of the mean.

compressive optic neuropathy, whereby a compressive lesion anywhere along the optic nerve or anterior visual pathway (anterior to the lateral geniculate body) causes axons to die, resulting in optic atrophy/pallor.[59] The ability to quantify sectoral pallor may provide additional value. For example, compression to the optic chiasm can cause pallor in the temporal and nasal zones—a condition known as band or bow tie atrophy.[60] Therefore, the pattern of optic disc pallor may further help localization of the lesion. Of particular relevance may be the detection of optic pathway gliomas (OPGs), which predominantly affect children (mean age at presentation, 8.8 years).[61] Assessment of vision is crucial in diagnosis; however, young children
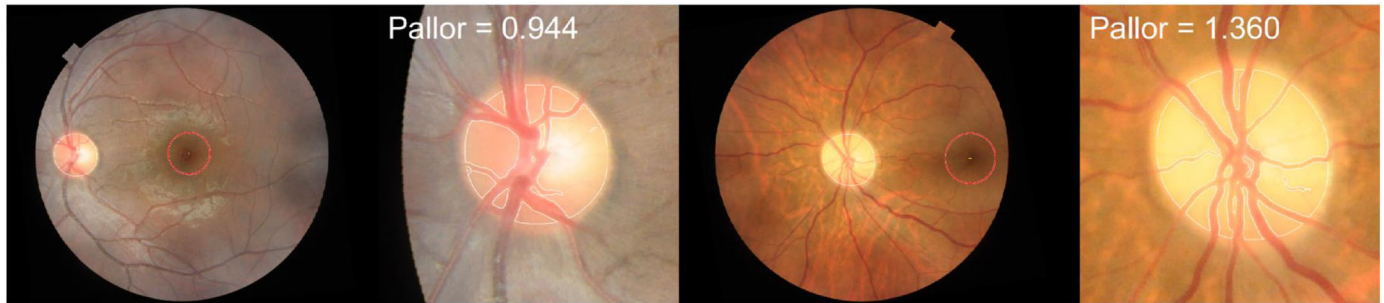
Figure 17. Intermittent stages of the pallor software on two images from the RFMiD dataset, one diagnosed by two ophthalmologists as having optic disc pallor (*right*) and a healthy control (*left*).

Table 6. Technical Characteristics of the Datasets Used to Assess How the Software Deals With Images Captured With a Range of Different Camera Systems, Resolutions, and Formats

Dataset	Format	Camera	Resolution	FOV	Centering	Dilation	Computation Time per Image (Seconds)
G1020	JPG	Topcon TRC-NW8	Between 1944 × 2108 and 2426 × 3007	45°	Mixed	Yes	2.8
MESSIDOR	TIFF	Topcon TRC-NW6	2240 × 1488	45°	Mixed	Yes	2.9
PREVENT (follow-up images)	BMP	Canon CR-Dgi	3072 × 2048	45°	Posterior pole	No	2.7
REFUGE (Canon)	JPG	Canon CR-2	1634 × 1634	–	Posterior pole	–	2.5
REFUGE (Zeiss)	JPG	Zeiss Visucam 500	2124 × 2056	–	Posterior pole	–	2.6

FOV, field of view.

will often not complain of vision loss, and instead present at a later stage with headache or pain.[62] Given that disc pallor is present in approximately 60% of cases,[61,62] it is feasible that OPG could be detected automatically through routine fundus imaging, which is not typically viewed by an ophthalmologist. Further research should investigate the association between optic disc pallor, as measured with the current software, and various types of compressive optic neuropathy.

Disc pallor is also an important measure of chemotherapy success in pediatric OPGs,[63] with the authors suggesting that the degree of pallor could be important. Indeed, complementary work investigating the visual outcomes of childhood OPG treated with radiotherapy found that severe disc pallor (compared with mild) at diagnosis or follow-up may be associated with a negative prognosis.[64] Further work could reeval-

uate such existing studies, substituting clinical notes on pallor for the continuous measures generated by our software (depending on the availability of fundus images). However, care should be taken if investigating changes in pallor, because it rarely improves[63]; therefore, the direction of change will almost always be one way.

OCT is the gold standard for assessing RNFL loss. However, compared with fundus, imaging it is costly, requires greater operator training, and is less prevalent. Furthermore, to avoid movement artefacts to which OCT is prone, such as those cause by ocular saccades, blinks, changes in head position, or respiratory movements,[65] patients must maintain a steady focus on a fixed point for several tens of seconds. Therefore, obtaining a high-quality OCT can be particularly challenging in individuals who may struggle with prolonged focus and steadiness, such as children,[66] the
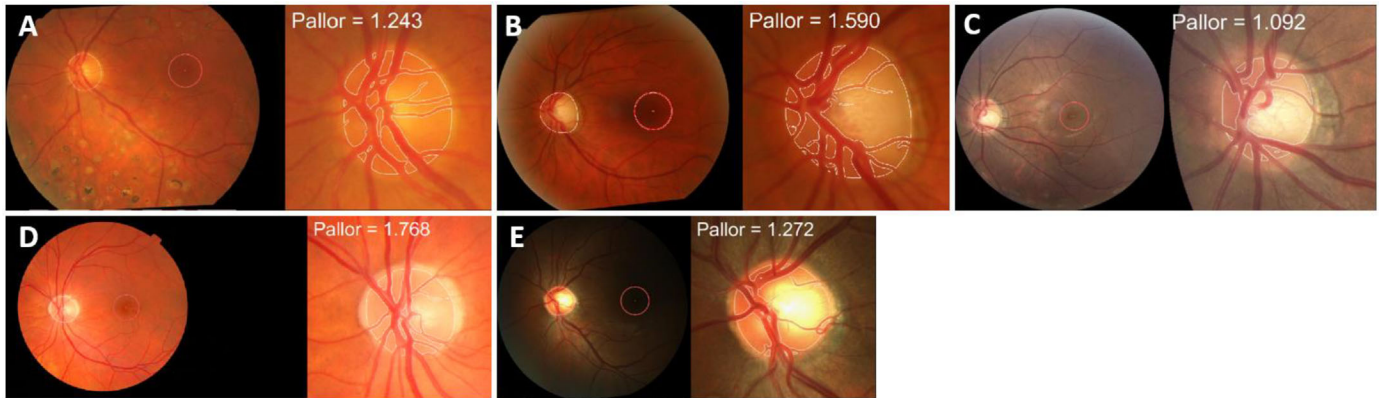
Figure 18. Robustness to camera system, format, and resolution results. (**A**) G1020, (**B**) MESSIDOR, (**C**) REFUGE Canon, (**D**) MESSIDOR, and (**E**) REFUGE Zeiss. Image to the left shows disc segmentation and fovea localization in the whole image, image to the right shows disc segmentation in closer detail. Pallor value shown is for the whole disc.
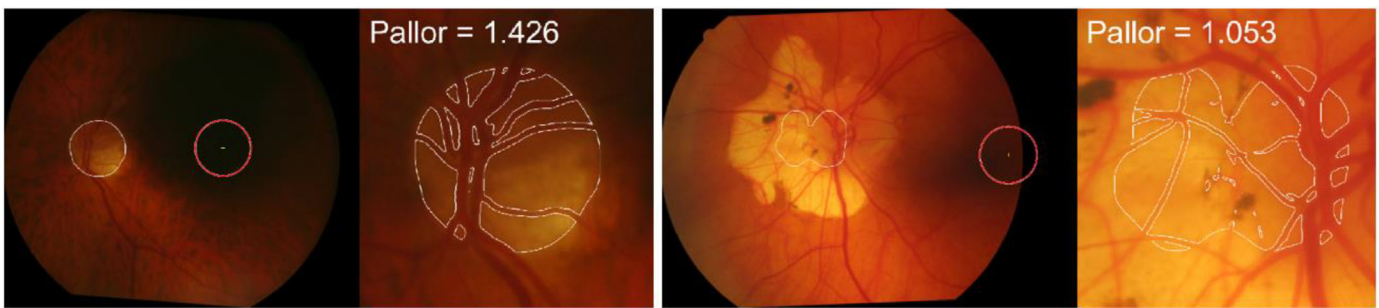

Figure 19. Automatic image rejection based on exceeding set thresholds for luminance (left) and eccentricity (right). We acknowledge that the image to the right was a failure of the software to correctly identify the disc margin owing to excessive chorioretinal atrophy.

Table 7. Automatic Image Rejection Thresholds

Criterion	Threshold	LBC		PREVENT	
		Mean $\pm$ SD	Images Over Threshold (*n*, %)	Mean $\pm$ SD	Images Over Threshold (*n*, %)
Eccentricity	>0.65	0.39 $\pm$ 0.11	22 (1.4)	0.4 $\pm$ 0.1	2 (0.6)
Luminance	<50	110.21 $\pm$ 26.24	18 (1.15)	90.56 $\pm$ 18.62	9 (2.85)

SD, standard deviation.

frail elderly, or those with movement disorders. Owing to the speed of acquisition, fundus imaging is more likely to be successful in these groups. Pallor derived from fundus photographs could provide an indicator that further examination is required, and could be a good alternative in groups where OCT scanning is not feasible.

Strengths of the current study included the networks' ability to segment the optic disc to the inner edge of the border tissue accurately. The disc margin is marked differently depending on the imaging modality through which it is observed. In OCT, the margin is marked at Bruch's membrane opening,[20] whereas in clinical ophthalmoscopy and fundus photography, it is defined as the inner edge of the border tissue.[20] There are many deep learning–based disc segmentation algorithms (for a review see Hasan et al.[28]); however, most systems are trained on one or more of four open-source image sets, namely IDRiD, RIMONE, DRISHTI-GS,[67] and DRIVE.[68] This factor may be problematic for our requirement, because, on close inspection of ground truth segmen-
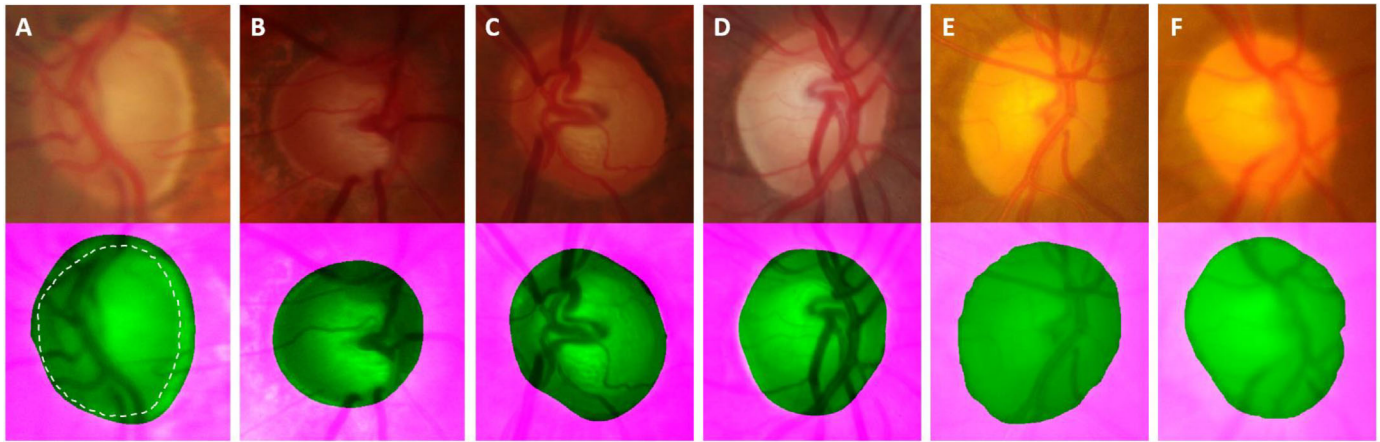
Figure 20. Examples from RIM ONE (glaucoma = **A**–**B**, nonglaucoma = **C**–**D**) and IDRiD (**E**–**F**), where the disc margin is overestimated according to a clinical definition. The ground truth (according to the original annotations) is marked in pink. The dashed line in (**A**) represents where we perceive the margin to be. In (**A**), the space between the dashed line and the start of the pink represents label noise.

tations in IDRiD and RIMONE, we observed a noticeable departure from what we perceived to be the clinically defined margin (Fig. 20). We believe this could be the result of averaging multiple annotations from different individuals to arrive at a ground truth, in which there is considerable disagreement. Such disagreements may have arisen owing to the annotators either marking the boundary clinically, or inferring Bruch's membrane opening–based information, for example, the bend of vessels at the rim. Although disagreement between multiple annotators provides an important measure of confidence that must be considered when assessing an automatic system, averaging multiple annotations can lead to label noise.

A recent study on label noise in medical image segmentation demonstrated that, although state-of-the-art networks are somewhat robust to unbiased or random noise, they are sensitive to biased noise.[69] Indeed, we observed that RIMONE and IDRiD may contain biased noise, whereby the clinically defined disc boundary is overestimated systematically with respect to where we perceive the clinical margin to lie. It is, therefore, possible that deep learning–based models trained on these datasets will systematically overestimate the clinically defined boundary. The significance of this factor for automatic segmentation programs would depend on the application. For example, the VAMPIRE software,[70] which is concerned with obtaining vessel-based measurements, requires that the disc be estimated as a best fit ellipse, in which case the precise boundary is less important. However, our work required greater precision, because sharp changes in color at the border would erroneously affect pallor metrics. Another use case

that may benefit from a more precisely defined disc margin is measuring the cup to disc ratio/profile in glaucoma,[71] because an overestimated rim could erroneously widen the profile, leading to a false positive (missing glaucoma).

Another strength of the current study is the robustness of the fovea detection network, which, in our experiments, gives good estimates even for very challenging images (Fig. 9). Additionally, locating the fovea helps the software to determine which eye (left or right) is being processed and allows for accurate zone placement. Last, the study had the advantage of using mixed effects modelling, which enabled the use of data from both eyes, thereby increasing statistical power.

The current study has several limitations. First, disc appearance is affected by physiological factors, chiefly the media opacity of the lens, but also potentially the richness of the capillary net supplying the optic nerve. Lens status was not available for the patients included in our study; therefore, we could not distinguish the pallor caused by optic atrophy from pseudopallor (nonpathological paleness, most notably caused by cataract extraction).[2] Further validation work should be carried out to assess the extent to which the current pallor metrics are affected by worsening cataract and cataract removal. In the meantime, information on cataract status and other potential causes of pseudopallor should be included as covariates where possible, particularly with older individuals. With regard to perfusion, in future studies it would be interesting to examine the relationship between OCT angiography measures of vessel density and pallor measured using the software to determine whether density of the capillary net might be a confound-

ing factor in grading pallor, or if the software could potentially be useful for evaluating changes to the optic nerve head caused by reduced optic nerve head perfusion.

Another limitation is that pallor quantification was affected somewhat by the brightness of the control region (which itself is largely determined by pigmentation—a factor that varies among individuals. To overcome this issue, we added control region brightness as a covariate in all statistical models. Although this approach may be acceptable for research studies with multiple participants, for clinical insight into a single image, normative data with a wide range of relevant parameters would be required to determine the extent to which pigmentation affects the measure.

With regard to peripapillary atrophy, we selected the control region to include as much of the background retina as possible, while minimizing the inclusion of any atrophy. Further, we used median values when calculating the overall control region brightness, which helped to mitigate the effect of any included atrophy. However, if the control region of an image contains a significant portion of atrophy, the image should be rejected, because the pallor metrics may be unreliable (there were no such cases in the PREVENT dataset).

Another limitation is that three of the four networks (disc localization and segmentation, and fovea localization) were partly trained on PREVENT images. This overlap calls into question the generalizability of the software. However, we point to internal (test set) and external testing carried out on all three networks, which show excellent generalizability. Nonetheless, further research should aim to replicate the finding that disc pallor is associated with pRNFL thickness in a novel dataset.

A final limitation was the lack of association between pRNFL thickness and pallor in five of the seven zones. In similar work, but using clinical notes (e.g., pallor absent/present), Aleman et al.[5] observed that the association between pallor and pRNFL thickness is optimal only when significant thinning has occurred. This may help explain the lack of associations in five zones in the current study, as participants in the PREVENT cohort are relatively healthy; global mean pRNFL was thick (98.2 ± 8.3 μm) in comparison with normative data (90 μm in Whites[72] and 94 μm in a multiethnic cohort),[72] although thinner than in individuals from Ghana (102 μm).[73]

Further research should aim to replicate the current findings in a larger sample, generate normative data, and test for associations with cardiovascular risk factors and disease.

Conclusions

A pale disc indicates irreversible damage to the anterior visual pathway and is present in numerous diseases. We present an automatic, artificial intelligence–enabled method that is fast, easy to use, robust, and suitable for application to large datasets. We found associations between pallor and pRNFL thickness, suggesting that disc pallor derived from fundus photographs may act as a proxy for pRNFL. We think our method will be useful for the identification and monitoring of the progression of diseases characterized by disc pallor and optic atrophy, including glaucoma, compression, and potentially in neurodegenerative disorders.

Acknowledgments

References

1. Ahmad SS, Kanukollu VM. Optic Atrophy. In: *StatPearls*. St Petersburg, FL: StatPearls Publishing; 2022. Accessed August 29, 2022. http://www.ncbi.nlm.nih.gov/books/NBK559130/.

2. Osaguona VB. Differential diagnoses of the pale/white/atrophic disc. *Community Eye Health*. 2016;29(96):71–74..

3. Ahmad SS, Kanukollu VM. Optic atrophy. *Handbook of Pediatric Retinal OCT and the Eye-Brain Connection*. Ner York: Elsevier; 2022:292–295, doi:10.1016/B978-0-323-60984-5.00064-0.

4. O'Neill EC, Danesh-Meyer HV, Kong GXY, et al. Optic disc evaluation in optic neuropathies: the optic disc assessment project. *Ophthalmology*. 2011;118(5):964–970, doi:10.1016/j.ophtha.2010.09.002.

5. Aleman TS, Huang J, Garrity ST, et al. Relationship between optic nerve appearance and retinal nerve fiber layer thickness as explored with spec-

tral domain optical coherence tomography. *Transl Vis Sci Technol*. 2014;3(6):4, doi:10.1167/tvst.3.6.4.

6. Ramm L, Schwab B, Stodtmeister R, et al. Assessment of optic nerve head pallor in primary open-angle glaucoma patients and healthy subjects. *Curr Eye Res*. 2017;42(9):1313–1318, doi:10.1080/02713683.2017.1307415.

7. Vilser W, Nagel E, Seifert BU, Riemer T, Weisensee J, Hammer M. Quantitative assessment of optic nerve head pallor. *Physiol Meas*. 2008;29(4):451–457, doi:10.1088/0967-3334/29/4/003.

8. Assad A, Caprioli J. Digital image analysis of optic nerve head pallor as a diagnostic test for early glaucoma. *Graefes Arch Clin Exp Ophthalmol*. 1992;230(5):432–436, doi:10.1007/BF00175928.

9. Nakano E, Hata M, Oishi A, et al. Quantitative comparison of disc rim color in optic nerve atrophy of compressive optic neuropathy and glaucomatous optic neuropathy. *Graefes Arch Clin Exp Ophthalmol*. 2016;254(8):1609–1616, doi:10.1007/s00417-016-3366-2.

10. Kang S, Kim US. Using ImageJ to evaluate optic disc pallor in traumatic optic neuropathy. *Korean J Ophthalmol KJO*. 2014;28(2):164–169, doi:10.3341/kjo.2014.28.2.164.

11. Ritchie CW, Ritchie K. The PREVENT study: a prospective cohort study to identify midlife biomarkers of late-onset Alzheimer's disease. *BMJ Open*. 2012;2(6):e001893, doi:10.1136/bmjopen-2012-001893.

12. Ritchie CW, Ritchie K. The PREVENT study: a prospective cohort study to identify midlife biomarkers of late-onset Alzheimer's disease. *BMJ Open*. 2012;2(6):e001893, doi:10.1136/bmjopen-2012-001893.

13. Taylor AM, Pattie A, Deary IJ. Cohort profile update: the Lothian birth cohorts of 1921 and 1936. *Int J Epidemiol*. 2018;47(4):1042–1060, doi:10.1093/ije/dyy022.

14. Zhang Z, Yin FS, Liu J, et al. ORIGA(-light): an online retinal fundus image database for glaucoma analysis and research. *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, Buenos Aires, Argentina. 2010;2010:3065–3068, doi:10.1109/IEMBS.2010.5626137.

15. Jin K, Huang X, Zhou J, et al. FIVES: a fundus image dataset for artificial intelligence based vessel segmentation. *Sci Data*. 2022;9(1):475, doi:10.1038/s41597-022-01564-3.

16. Fumero F, Alayon S, Sanchez JL, Sigut J, Gonzalez-Hernandez M. RIM-ONE: an open retinal image database for optic nerve evaluation. In: *2011 24th International Symposium on Computer-Based Medical Systems (CBMS)*. 2011:1–6, doi:10.1109/CBMS.2011.5999143.

17. Porwal P, Pachade S, Kamble R, et al. Indian Diabetic Retinopathy Image Dataset (IDRiD): a database for diabetic retinopathy screening research. *Data*. 2018;3(3):25, doi:10.3390/data3030025.

18. Pachade S, Porwal P, Thulkar D, et al. Retinal fundus multi-disease image dataset (RFMID): A dataset for multi-disease detection research. *Data*. 2021;6(2):1–14, doi:10.3390/data6020014.

19. Strouthidis N, Yang H, Reynaud J, et al. Comparison of clinical and spectral domain optical coherence tomography optic disc margin anatomy. *Invest Ophthalmol Vis Sci*. 2009;50(10):4709–4718, doi:10.1167/iovs.09-3586.

20. Chauhan BC, Burgoyne CF. From clinical examination of the optic disc to clinical assessment of the optic nerve head: a paradigm change. *Am J Ophthalmol*. 2013;156(2):218–227.e2, doi:10.1016/j.ajo.2013.04.016.

21. Yang F, Zamzmi G, Angara S, et al. Assessing inter-annotator agreement for medical image segmentation. *IEEE Access Pract Innov Open Solut*. 2023;11:21300–21312, doi:10.1109/access.2023.3249759.

22. Minaee S, Boykov Y, Porikli F, Plaza A, Kehtarnavaz N, Terzopoulos D. Image segmentation using deep learning: a survey. *IEEE Trans Pattern Anal Mach Intell*. 2022;44(7):3523–3542, doi:10.1109/TPAMI.2021.3059968.

23. Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with Atrous separable convolution for semantic image segmentation. *Proceedings of the European Conference on Computer Vision (ECCV)*. *arXiv*. 2018;1802.02611. Published online August 22, 2018, doi:10.48550/arXiv.1802.02611.

24. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: inverted residuals and linear bottlenecks. *arXiv*. 2019;1801.04381 [cs.CV]. Published online March 21, 2019, doi:10.48550/arXiv.1801.04381.

25. Chollet F. Xception: deep learning with depthwise separable convolutions. *arXiv*. 2017;1610.02357 [cs.CV]. Published online April 4, 2017, doi:10.48550/arXiv.1610.02357.

26. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009:248–255, doi:10.1109/CVPR.2009.5206848.

27. Yu S, Xiao D, Frost S, Kanagasingam Y. Robust optic disc and cup segmentation with deep learning for glaucoma detection. *Comput Med Imaging Graph*. 2019;74:61–71, doi:10.1016/j.compmedimag.2019.02.005.

28. Hasan MdK, Alam MdA, Elahi MdTE, Roy S, Martí R. DRNet: Segmentation and localization of optic disc and Fovea from diabetic retinopathy image. *Artif Intell Med*. 2021;111:102001, doi:10.1016/j.artmed.2020.102001.

29. Al-Bander B, Al-Nuaimy W, Williams BM, Zheng Y. Multiscale sequential convolutional neural networks for simultaneous detection of fovea and optic disc. *Biomed Signal Process Control*. 2018;40:91–101, doi:10.1016/j.bspc.2017.09.008.

30. Yang HK, Oh JE, Han SB, Kim KG, Hwang JM. Automatic computer-aided analysis of optic disc pallor in fundus photographs. *Acta Ophthalmol (Copenh)*. 2019;97(4):e519–e525, doi:10.1111/aos.13970.

31. Gonzalez-Hernandez M, Gonzalez-Hernandez D, Perez-Barbudo D, Rodriguez-Esteve P, Betancor-Caro N, de la Rosa MG. Fully automated colorimetric analysis of the optic nerve aided by deep learning and its association with perimetry and oct for the study of glaucoma. *J Clin Med*. 2021;10(15):3231, doi:10.3390/jcm10153231.

32. MacGillivray TJ, Cameron JR, Zhang Q, et al. Suitability of UK Biobank Retinal Images for Automatic Analysis of Morphometric Properties of the Vasculature. *PLoS One*. 2015;10(5):e0127914, doi:10.1371/JOURNAL.PONE.0127914.

33. Shuang YG, Maguire MG, Glynn R, Rosner B. Tutorial on biostatistics: linear regression analysis of continuous correlated eye data. *Ophthalmic Epidemiol*. 2017;24(2):130–140, doi:10.1080/09286586.2016.1259636.

34. VanderWeele TJ. Principles of confounder selection. *Eur J Epidemiol*. 2019;34(3):211–219, doi:10.1007/s10654-019-00494-6.

35. Berendschot TTJM, DeLint PJ, van Norren D. Fundus reflectance—historical and present ideas. *Prog Retin Eye Res*. 2003;22(2):171–200, doi:10.1016/S1350-9462(02)00060-5.

36. Bajwa MN, Singh GAP, Neumeier W, Malik MI, Dengel A, Ahmed S. G1020: a benchmark retinal fundus image dataset for computer-aided glaucoma detection. *arXiv*. 2020;2006.09158 [eess.IV]. Published online May 28, 2020, doi:10.48550/arXiv.2006.09158.

37. Decencière E, Zhang X, Cazuguel G, et al. Feedback on a publicly distributed image database: the messidor database. *Image Anal Stereol*. 2014;33(3):231–234, doi:10.5566/ias.1155.

38. Orlando JI, Fu H, Breda JB, et al. REFUGE challenge: a unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Med Image Anal*. 2020;59:101570, doi:10.1016/j.media.2019.101570.

39. Camejo L, Noecker RJ. CHAPTER 14 - Optic nerve imaging. In: Stamper RL, Lieberman MF, Drake MV, eds. *Becker-Shaffer's Diagnosis and Therapy of the Glaucomas* (Eighth Edition). St. Louis: Mosby; 2009:171–187, doi:10.1016/B978-0-323-02394-8.00014-0.

40. Wong TY, Mitchell P. Hypertensive retinopathy. *N Engl J Med*. 2004;351(22):2310–2317, doi:10.1056/NEJMra032865.

41. Abu El-Asrar AM, Herbort CP, Tabbara KF. Differential diagnosis of retinal vasculitis. *Middle East Afr J Ophthalmol*. 2009;16(4):202–218, doi:10.4103/0974-9233.58423.

42. Kumar J, Wierzbicki AS. Lipemia retinalis. *N Engl J Med*. 2005;353(8):823–823, doi:10.1056/NEJMicm040437.

43. Fu H, Cheng J, Xu Y, Wong DWK, Liu J, Cao X. Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. *IEEE Trans Med Imaging*. 2018;37(7):1597–1605, doi:10.1109/TMI.2018.2791488.

44. Sevastopolsky A. Optic disc and cup segmentation methods for glaucoma detection with modification of U-Net convolutional neural network. *Pattern Recognit Image Anal*. 2017;27(3):618–624, doi:10.1134/S1054661817030269.

45. Meng Y, Zhang H, Zhao Y, et al. Graph-based region and boundary aggregation for biomedical image segmentation. *IEEE Trans Med Imaging*. 2022;41(3):690–701, doi:10.1109/TMI.2021.3123567.

46. Leung CKS, Choi N, Weinreb RN, et al. Retinal nerve fiber layer imaging with spectral-domain optical coherence tomography: pattern of RNFL defects in glaucoma. *Ophthalmology*. 2010;117(12):2337–2344, doi:10.1016/j.ophtha.2010.04.002.

47. Chen Y, Yuan Y, Zhang S, et al. Retinal nerve fiber layer thinning as a novel fingerprint for cardiovascular events: results from the prospective cohorts in UK and China. *BMC Med*. 2023;21(1):24, doi:10.1186/s12916-023-02728-7.

48. Thomson KL, Yeo JM, Waddell B, Cameron JR, Pal S. A systematic review and meta-analysis of retinal nerve fiber layer change in dementia, using optical coherence tomography. *Alzheimers*

Dement Diagn Assess Dis Monit. 2015;1(2):136–143, doi:10.1016/j.dadm.2015.03.001.

49. Ko F, Muthy ZA, Gallacher J, et al. Association of retinal nerve fiber layer thinning with current and future cognitive decline: a study using optical coherence tomography. *JAMA Neurol*. 2018;75(10):1198–1205, doi:10.1001/jamaneurol.2018.1578.

50. Mutlu U, Colijn JM, Ikram MA, et al. Association of retinal neurodegeneration on optical coherence tomography with dementia: a population-based study. *JAMA Neurol*. 2018;75(10):1256–1263, doi:10.1001/jamaneurol.2018.1563.

51. Biffi E, Turple Z, Chung J, Biffi A. Retinal biomarkers of cerebral small vessel disease: a systematic review. *PLoS One*. 2022;17(4):e0266974..

52. Wang D, Li Y, Wang C, et al. Localized retinal nerve fiber layer defects and stroke. *Stroke*. 2014;45(6):1651–1656, doi:10.1161/STROKEAHA.113.004629.

53. Conn FL. When glaucomatous damage isn't glaucoma. Accessed April 17, 2023. https://www.reviewofophthalmology.com/article/when-glaucomatous-damage-isnt-glaucoma.

54. Hwang YH, Song M, Kim YY, Yeom DJ, Lee JH. Interocular symmetry of retinal nerve fibre layer thickness in healthy eyes: a spectral-domain optical coherence tomographic study. *Clin Exp Optom*. 2014;97(6):550–554, doi:10.1111/cxo.12218.

55. Dalgliesh JD, Tariq YM, Burlutsky G, Mitchell P. Symmetry of retinal parameters measured by spectral-domain OCT in normal young adults. *J Glaucoma*. 2015;24(1):20, doi:10.1097/IJG.0b013e318287ac2f.

56. Yang M, Wang W, Xu Q, Tan S, Wei S. Interocular symmetry of the peripapillary choroidal thickness and retinal nerve fibre layer thickness in healthy adults with isometropia. *BMC Ophthalmol*. 2016;16(1):182, doi:10.1186/s12886-016-0361-7.

57. Budenz DL. Symmetry between the right and left eyes of the normal retinal nerve fiber layer measured with optical coherence tomography (an AOS thesis). *Trans Am Ophthalmol Soc*. 2008;106:252–275..

58. Cameron JR, Megaw RD, Tatham AJ, et al. Lateral thinking – interocular symmetry and asymmetry in neurovascular patterning, in health and disease. *Prog Retin Eye Res*. 2017;59:131–157, doi:10.1016/j.preteyeres.2017.04.003.

59. Rodriguez-Beato FY, De Jesus O. Compressive optic neuropathy. In: *StatPearls*. St Petersburg, FL: StatPearls Publishing; 2023. Accessed April 17, 2023. http://www.ncbi.nlm.nih.gov/books/NBK560583/.

60. Monteiro MLR. Optical coherence tomography analysis of axonal loss in band atrophy of the optic nerve. *Br J Ophthalmol*. 2004;88(7):896–899, doi:10.1136/bjo.2003.038489.

61. Fried I, Tabori U, Tihan T, Reginald A, Bouffet E. Optic pathway gliomas: a review. *CNS Oncol*. 2013;2(2):143–159, doi:10.2217/cns.12.47.

62. Huang M, Patel J, Patel BC. Optic Nerve Glioma. In: *StatPearls*. St Petersburg, FL: StatPearls Publishing; 2023. Accessed April 18, 2023. http://www.ncbi.nlm.nih.gov/books/NBK557878/.

63. Fisher MJ, Loguidice M, Gutmann DH, et al. Visual outcomes in children with neurofibromatosis type 1–associated optic pathway glioma following chemotherapy: a multicenter retrospective analysis. *Neuro-Oncol*. 2012;14(6):790–797, doi:10.1093/neuonc/nos076.

64. Campagna M, Opocher E, Viscardi E, et al. Optic pathway glioma: long-term visual outcome in children without neurofibromatosis type-1. *Pediatr Blood Cancer*. 2010;55(6):1083–1088, doi:10.1002/pbc.22748.

65. Chhablani J, Krishnan T, Sethi V, Kozak I. Artifacts in optical coherence tomography. *Saudi J Ophthalmol*. 2014;28(2):81–87, doi:10.1016/j.sjopt.2014.02.010.

66. Lee H, Proudlock FA, Gottlob I. Pediatric optical coherence tomography in clinical practice—recent progress. *Invest Ophthalmol Vis Sci*. 2016;57(9):OCT69–OCT79, doi:10.1167/iovs.15-18825.

67. Sivaswamy J, Krishnadas SR, Datt Joshi G, Jain M, Syed Tabish AU. Drishti-GS: retinal image dataset for optic nerve head (ONH) segmentation. In: *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, Beijing, China. 2014:53–56, doi:10.1109/ISBI.2014.6867807.

68. Staal J, Abramoff MD, Niemeijer M, Viergever MA, van Ginneken B. Ridge-based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging*. 2024;23(4):501–509, doi:10.1109/TMI.2004.825627.

69. Vorontsov E, Kadoury S. Label Noise in Segmentation Networks: Mitigation Must Deal with Bias. In: Engelhardt S, Oksuz I, Zhu D, et al., Eds. *Deep Generative Models, and Data Augmentation, Labelling, and Imperfections*. Lecture Notes in Computer Science. New York: Springer International Publishing; 2021:251–258, doi:10.1007/978-3-030-88210-5_25.

70. Mookiah MRK, Hogg S, MacGillivray T, Trucco E. On the quantitative effects of compression of retinal fundus images on morphometric vascular measurements in VAMPIRE. *Comput Methods*

Programs Biomed. 2021;202:105969, doi:10.1016/j.cmpb.2021.105969.

71. MacCormick IJC, Williams BM, Zheng Y, et al. Correction: Accurate, fast, data efficient and interpretable glaucoma diagnosis with automated spatial analysis of the whole cup to disc profile (PLoS One (2019) 14:1 (e0209409) DOI: 10.1371/journal.pone.0209409). *PLoS One*. 2019;14(4):1–20, doi:10.1371/journal.pone.0215056.

72. Knight OJ, Girkin CA, Budenz DL, Durbin MK, Feuer WJ, Cirrus OCT Normative Database Study Group. Effect of race, age, and axial length on optic nerve head parameters and retinal nerve fiber layer thickness measured by Cirrus HD-OCT. *Arch Ophthalmol*. 2012;130(3):312–318, doi:10.1001/archopthalmol.2011.1576.

73. Ocansey S, Abu EK, Owusu-Ansah A, et al. Normative values of retinal nerve fibre layer thickness and optic nerve head parameters and their association with visual function in an African population. *J Ophthalmol*. 2020;2020:e7150673, doi:10.1155/2020/7150673.