Supplementary Material to

A functional regression model of the retinal nerve fiber layer thickness in healthy subjects

Ivania Pereira, Eleonore Pablik, Florian Schwarzhans, Hemma Resch, Georg Fischer, Clemens Vass, Florian Frommlet

Further details on Statistical Methods

1) Landmark registration

Landmark registration was performed by aligning individual RNFL curves according to three prominent registration points illustrated in Figure S1. We thereby introduce a new coordinate system (denoted by x) in which RNFL curves of all subjects are better comparable which immediately reduces the intersubject variability.

Example of the registration points M, P1, and P2 in the RNFL curve of one subject

Figure S1. Typical RNFL curve over the 256 sectors which illustrates the three registration points used for alignment of individual curves. M is the minimum of the curve in the vicinity of the initial sector (which due to periodicity may also be close to sector 256). P1 and P2 are the peaks on the right hand side and the left hand side of M, respectively, which are typically found in the neighborhoods of sector 59 and sector 212. In the new coordinate system the registration points M, P1, and P2 will be located at x_M=0, x_P1=59, and x_P2=212.
Each healthy subject has a characteristic minimum $M$ near the fovea (sector 0), which was defined as the origin of the new aligned coordinate system. Due to periodicity of the RNFL thickness profile, the minimum value will appear again after one period. We kept the period length of 256, which corresponds to the number of sectors from the original data. The minimum $M$ is flanked by two characteristic maximum peaks, defined as $P_1$ and $P_2$, one located in the superior and the other in the inferior area of the eye. These peaks were found in the RNFL profile of each subject by local search algorithms, where in our training set the median position of $P_1$ was located at sector 59 and the median position of $P_2$ at sector 212.

For each individual $i$ a new RNFL profile $\tilde{Y}_i(x)$ was computed by piecewise linearly rescaling the coordinate system such that the first peak was shifted to $x_{P_1} = 59$ and the second peak to $x_{P_2} = 212$. By linear interpolation the shifted curves were then evaluated on the integers 0 to 255 in order to have identical $x$ values for all subjects for the estimation of our coefficients. Figure S2 illustrates the interpolation function for all 202 individuals of the training and validation sample. The variation of landmark positions is not particularly dramatic and more sophisticated interpolation approaches are not needed.

![Figure S2. Function of piecewise linear rescaling (time-warping function) for 202 individuals (training and validation sample) used after aligning with respect to the minimum M.](http://tvst.arvojournals.org/pdfaccess.ashx?url=/data/journals/tvst/936672/)
2) Functional principal component analysis

Next functional principal components analysis (fPCA) was applied on the aligned RNFL curves. Due to the inherent smoothness of the RNFL curves fPCA was applied directly and no smoothing by Fourier basis methods was applied before fPCA. The first 16 principal components $\Psi_j$, $j = 1,\ldots,16$, accounted for 97.56% of the observed inter subject variance (see Figure S3). These were used in the second step of our functional approach as the 16 basis functions to approximately model the aligned RNFL curves $\tilde{Y}_i(x)$ according to the model

$$\tilde{Y}_i(x) = \tilde{Y}_0(x) + \sum_{j=1}^{16} B_{ij} \times \Psi_j(x) + \epsilon_i(x) \quad (1)$$

Here $\tilde{Y}_0(x)$ represents the average of peak adjusted RNFL curves of healthy subjects and will be estimated by taking the average over our training sample. The coefficients $B_{ij}$ are computed by the method of least squares. Specifically, we are minimizing the squared sum of error terms

$$\sum_{k=0}^{255} \left( \tilde{Y}_i(k) - \tilde{Y}_0(k) - B_{ij} \times \Psi_j(k) \right)^2, \quad (2)$$

where the values of $\tilde{Y}_i(k)$ are obtained by linear interpolation from the peak adjusted RNFL curves. The estimates of the average values were then obtained as

$$\tilde{Y}_0(k) = \frac{1}{n} \sum_{i=0}^{n} \tilde{Y}_i(k) \quad (3).$$

Figure S3. Scree plot of the first 32 eigenvalues from fPCA (left panel) and the corresponding percentage of explained variation.
Clearly, the OCT measurements of the RNFL thickness profiles are themselves not completely exact. Based on additional data from 73 healthy subjects with repeated OCT measurements we can estimate the magnitude of the error of OCT measurements. We are interested both in the maximum of absolute deviation over all 256 sectors ($L_\infty$ norm) and the RMSE over all sectors which corresponds to the $L_2$ norm. Figure S4 illustrates the comparison of these OCT errors with fPCA errors like $\epsilon_i(x)$ in equation (1) but using different numbers of principal components. It is apparent that using 12 components fPCA errors become comparable in magnitude to the OCT measurement error and using 16 components the fPCA errors actually get smaller than OCT measurement errors. RMSE between two measurements of one patient was on average 5.23µm, while the average root mean square of the observed $\epsilon_i(x)$ in our model was 2.69µm only.

Figure S4. Comparison of OCT measurement errors in 73 healthy subjects with repeated measurements and fPCA model errors using different numbers of principal components ranging from 4 to 20. $L_\infty$ refers to the maximal absolute difference over all 256 sectors whereas $L_2$ corresponds to the root mean squared error RMSE.
Note that at this point we have reduced the number of parameters to model the RNFL profile of each subject to 16 PCA loadings $B_{ij}$ plus three shift parameters from the alignment procedure of the first step. The alignment step was already designed to reduce variability between subjects, whereas the second step was designed to model the RNFL profile curves with a fairly small number of parameters.

3) Regression model

In the third step, we will now consider regression models of the parameters $B_{ij}$ using specific information of individuals to further reduce the variability between subjects. The regressors consist of vessel thickness and the 7 subject specific parameters listed in the main manuscript. However, with respect to vessel thickness we did not use a smoothed RVD curve like Pereira et al.\textsuperscript{6}, but instead considered for each $X_{ir}$ the original blood vessel data at sector $r$ and computed for each basis function a weighted average of thickness,

$$V_{ij} = \sum_r \Psi_j(x_r) X_{ir} \quad (4)$$

The variable $V_{ij}$ quantifies the amount of blood vessel thickness associated with the $j$-th principal component. Because the RNFL thickness profile was modelled using the same principal components as basis functions we introduced the following regression models

$$B_{ij} = V_{ij} \beta^{(j)}_v + \sum_{l=1}^7 C_{ij} \beta^{(j)}_l, \quad j = 1, \ldots, 16. \quad (5)$$

The first regressor $V_{ij}$ modeled the effect of vessel thickness specifically associated with the $j$-th principal component and $C_{ij}$ referred to the 7 additional subject specific characteristics described above. There are now 128 parameters to be estimated. To reduce the potential number of regressors for our final model, we performed backward model selection separately for each $j$ using the Akaike Information Criterion (AIC). Local vessel thickness $V_{ij}$ remained in all 16 regression models, while for the 7 remaining covariates only 33 out of 112 coefficients were selected to be non-zero. Thus, for our final model, we have to estimate 49 parameters.