Focused Tortuosity Definitions Based on Expert Clinical Assessment of Corneal Subbasal Nerves

Neil Lagali, Enea Poletti, Dipika V. Patel, Charles N. J. McGhee, Pedram Hamrah, Ahmad Kheirkhah, Mitra Tavakoli, Ioannis N. Petropoulos, Rayaz A. Malik, Tor Paaske Utheim, Andrey Zhivov, Oliver Stachs, Karen Falke, Sabine Peschel, Rudolf Guthoff, Cecilia Chao, Blanka Golebiowski, Fiona Stapleton and Alfredo Ruggeri

Linköping University Post Print

N.B.: When citing this work, cite the original article.

Original Publication:

Neil Lagali, Enea Poletti, Dipika V. Patel, Charles N. J. McGhee, Pedram Hamrah, Ahmad Kheirkhah, Mitra Tavakoli, Ioannis N. Petropoulos, Rayaz A. Malik, Tor Paaske Utheim, Andrey Zhivov, Oliver Stachs, Karen Falke, Sabine Peschel, Rudolf Guthoff, Cecilia Chao, Blanka Golebiowski, Fiona Stapleton and Alfredo Ruggeri, Focused Tortuosity Definitions Based on Expert Clinical Assessment of Corneal Subbasal Nerves, 2015, Investigative Ophthalmology and Visual Science, (56), 9, 5102-5109.

<http://dx.doi.org/10.1167/iovs.15-17284>

Copyright: Association for Research in Vision and Ophthalmology (ARVO)

<http://www.arvo.org/>

Postprint available at: Linköping University Electronic Press

<http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-122543>

Focused Tortuosity Definitions Based on Expert Clinical Assessment of Corneal Subbasal Nerves

Neil Lagali^{1*}, Enea Poletti², Dipika V. Patel³, Charles N.J. McGhee³, Pedram Hamrah⁴, Ahmad Kheirkhah⁴, Mitra Tavakoli⁵, Ioannis N. Petropoulos^{5,6}, Rayaz A. Malik^{5,6}, Tor Paaske Uttheim^{7,8}, Andrey Zhyvov⁹, Oliver Stachs⁹, Karen Falke⁹, Sabine Peschel⁹, Rudolf Guthoff⁹, Cecilia Chao¹⁰, Blanka Golebiowski¹⁰, Fiona Stapleton¹⁰, Alfredo Ruggeri²

¹Department of Ophthalmology, Institute for Clinical and Experimental Medicine, Faculty of Health Sciences, Linköping University, 581 83 Linköping, Sweden

²Department of Information Engineering, University of Padova, via G.Gradenigo 6/B, 35131 Padova, Italy

³Department of Ophthalmology, New Zealand National Eye Centre, Faculty of Medical and Health Sciences, University of Auckland, Private Bag 92019, Auckland, New Zealand 1142

⁴Ocular Surface Imaging Center, Schepens Eye Research Institute, Department of Ophthalmology, Harvard Medical School, 243 Charles Street, Boston, MA 02114

⁵Centre for Endocrinology & Diabetes, Institute of Human Development, University of Manchester, 46 Grafton Street, Manchester M13 9NT, UK

⁶Weill Cornell Medical College in Qatar, Qatar Foundation, Education City, PO Box: 24144-Doha, Qatar

⁷Department of Medical Biochemistry, Oslo University Hospital, Norway

⁸Department of Oral Biology, Faculty of Dentistry, University of Oslo, Oslo, Norway

⁹Department of Ophthalmology, University of Rostock, Doberaner Str. 140, 18057 Rostock, Germany

¹⁰School of Optometry and Vision Science, Level 3, North Wing, RMB, Gate 14, Barker Street, The University of New South Wales, Sydney NSW 2052, Australia

*corresponding author

Running title: Expert assessment of corneal nerve tortuosity

Word count: Abstract 249, Manuscript excluding abstract 3206

Address for correspondence:

Neil Lagali
Department of Ophthalmology,
Institute for Clinical and Experimental Medicine
Faculty of Health Sciences
Linköping University
581 83 Linköping, Sweden
Email: neil.lagali@liu.se
Tel: +46 10 1034658
Fax: +46 10 1033065

ABSTRACT

Purpose: To examine agreement among experts in the assessment of corneal subbasal nerve tortuosity.

Methods: Images of corneal subbasal nerves were obtained from investigators at seven sites (Auckland, Boston, Linköping, Manchester, Oslo, Rostock, and Sydney) using laser-scanning in vivo confocal microscopy. A set of 30 images was assembled and ordered by increasing tortuosity by ten expert graders from the seven sites. In a first experiment, graders assessed tortuosity without a specific definition and performed grading three times, with at least one week between sessions. In a second experiment, graders assessed the same image set using four focused tortuosity definitions. Inter-session and inter-grader repeatability for the experiments were determined using the Spearman rank correlation.

Results: Expert graders without a specific tortuosity definition had high inter-session (Spearman correlation coefficient 0.80) but poor inter-grader (0.62) repeatability. Specific definitions improved inter-grader repeatability to 0.79. In particular, tortuosity defined by frequent small-amplitude directional changes (short range tortuosity) or by infrequent large-amplitude directional changes (long range tortuosity), indicated largely independent measures and resulted in improved repeatability across the graders. A further refinement – grading only the most tortuous nerve in a given image – improved the average correlation of a given grader's ordering of images with the group average to 0.86 – 0.90.

Conclusions: Definitions of tortuosity specifying short or long-range tortuosity and considering only the most tortuous nerve in an image improved the agreement in tortuosity grading among a group of expert observers. These definitions could improve accuracy and consistency in quantifying subbasal nerve tortuosity in clinical studies.

Introduction

Tortuosity is defined by the degree of twistedness of a curved structure. It is a particularly evident feature of corneal subbasal nerves. Tortuosity is becoming more widely used to quantitatively describe and/or compare subbasal nerve status in healthy corneas,¹⁻⁵ in the context of ocular pathology such as dry eye disease⁵⁻⁷ and keratoconus,⁴ and in systemic diseases such as diabetes,⁸⁻¹⁵ rheumatoid arthritis,¹⁶ inflammatory polyneuropathy,¹⁷ Wilson's disease,¹⁸ and amyotrophic lateral sclerosis.¹⁹ Unfortunately, however, there is no standardized definition for tortuosity, and its interpretation and quantification can vary, leading to poor reproducibility for this measure.^{20,21} While some methods to describe tortuosity are purely quantitative and mathematically-defined,⁴ others are semi-quantitative and rely on observer judgment,¹ and still others are based on quantitative descriptions derived from subjective observation.⁸ The plurality of definitions not only makes it difficult to compare results across studies, but may also result in overlooking important differences in nerve morphology by adopting a particular definition. Computation of a single value for tortuosity may for example mask the complexity of nerve patterns, such that widely differing morphologic characteristics of nerves could have the same numeric tortuosity value.

As a starting point for developing more standardized and reproducible measures of tortuosity, the interpretation and perception of nerve tortuosity is important. Human judgement may take into account nerve architecture and morphologic features that have not yet been explicitly identified or mathematically defined. Also, an ideal tortuosity grading and quantification scheme should be consistent with the judgment of those experienced in assessing nerve images and making clinical decisions. To date, the perception of tortuosity and the level of agreement between expert observers have not been fully explored.

This study therefore had several aims:

1. to investigate the perception of nerve tortuosity and measure its intra- and inter-observer agreement across a range of subbasal nerve images from healthy and pathologic corneas, representing patients and experienced graders from geographically diverse regions;
2. to explore alternative, more focused definitions of tortuosity and determine the agreement of the perception of these focused definitions among a group of expert graders; and

3. to provide reference image sets representing ground-truth agreement among experienced graders – for the future development of new algorithms to measure tortuosity in clinical studies.

Starting with a reference set of diverse images and a group of examiners experienced in the assessment of corneal nerve images, it was hypothesized that a consensus would emerge, that could eventually lead to standardization in the assessment and quantification of subbasal nerve tortuosity.

Methods

Image compilation across centers

Investigators from seven geographically dispersed clinical research centers (Auckland, Boston, Linköping, Manchester, Oslo, Rostock, and Sydney) were each asked to contribute six to ten images of the subbasal nerve plexus depicting various severities of tortuous subbasal nerves (by subjective assessment). Images were selected from a database at each investigator clinic. All centers used the same type of laser-scanning in vivo confocal microscope to acquire the images (Heidelberg Retinal Tomograph 2/3 with Rostock Corneal Module, HRT-RCM, Heidelberg Engineering, Heidelberg, Germany), and all images were from patients who had given informed consent at the time of examination. All images were anonymized at the center of origin by removing identifying patient information except for age, gender, and disease. As the acquisition of original images from individual centers was approved by the respective local ethical review committees, occurred with informed consent, and followed the tenets of the Declaration of Helsinki, no specific further ethical approval was sought for a retrospective analysis of the resulting compilation of images. From the 67 received images, a set of 30 images was selected by including high-quality images without tilt (oblique orientation) and depicting clear, high-contrast subbasal nerves while excluding similar (redundant) images. The final images represented 30 eyes from 30 subjects with the following conditions: healthy (10 subjects), dry eye disease (4), post-LASIK (4), diabetes mellitus type 2 (3), diabetes mellitus type 1 (2), keratoconus (2), corneal transplantation (1), contact lens wearer (1), granular dystrophy (1), autism (1), and untreated gastrointestinal cancer (1).

124 **Experiment 1: Grading of generic tortuosity**

125

126 Ten experienced clinical graders from the seven centers were asked to sort the set of 30 images by
127 arranging them in order of increasing tortuosity. To facilitate this task, a custom-developed image
128 sorting tool was provided, TorTsorT (available for download at <http://bioimlab.dei.unipd.it>). TorTsorT
129 sequentially displays pairs of images, asking the user to select the more tortuous image. It then
130 combines all the sorted pairs into a sequence of images by means of the merge-sort algorithm.
131 Merge-sort is a comparison-based divide and conquer sorting algorithm that ensures the best
132 average- and worst-case performance. Using this software, each grader provided input to sort the set
133 of 30 images from lowest to highest tortuosity.

134 No specific definition of tortuosity or instruction on how to assess tortuosity was provided to the
135 graders. Graders performed the sorting three times (sessions) on the same image set. Sessions were
136 separated in time by at least one week, with images to be analyzed at a given session provided to the
137 centers with one-week delays after being randomly renamed, so as to minimize any memory effect.
138 Besides the sorting of images, graders were also asked to note any difficulties or comments they had
139 while performing the sorting during the multiple sessions.

140

141 **Experiment 2: Grading by specific definitions of tortuosity**

142

143 Based on grader comments in the first experiment and inspection of the images which had the
144 greatest average differences in position in the ordered list by the various graders, some specific
145 definitions of tortuosity were provided to the graders. Four different definitions of tortuosity were
146 proposed, abbreviated T1 - T4:

147

148 **T1:** a measure of short-range changes in direction, considering an average of all nerves in the image

149 **T2:** a measure of short-range changes in direction, considering only the nerve with maximum
150 tortuosity

151 **T3:** a measure of long-range changes in direction, considering an average of all nerves in the image

152 **T4:** a measure of long-range changes in direction, considering only the nerve with maximum
153 tortuosity

154

155 These short- and long-range definitions are illustrated in Figure 1. The ten graders were then asked to
 156 sort again the same 30 image set from Experiment 1 in order of increasing tortuosity, using each of
 157 the four definitions T1 – T4.

158

159

160 Figure 1. Schematic illustration of the more focused tortuosity definitions used in Experiment 2. Short-
 161 range directional changes were often small-amplitude and high frequency (short periodicity) while
 162 long-range directional changes tended to be large-amplitude and lower frequency (long periodicity).

163

164 **Consensus rank**

165

166 In order to express an overall or average image ordering, e.g., to derive the ground truth assessment
 167 to compare with and to enable comparison between graders and/or tortuosity definitions, we defined
 168 the *consensus rank*, which is the equivalent for ordered lists of the average for numerical quantities.
 169 For a given set of ordered lists of images, the *consensus rank* was determined as follows: each image
 170 was given a score represented by its average position (rank) in the ordered lists; the *consensus rank*
 171 was obtained by ordering the images by these scores.

172 In experiment 1, to assess inter-grader agreement, the Spearman rank correlation coefficient was
 173 calculated between the ordered list of images from a given grader at a given session and the
 174 *consensus rank* (see below) of all other graders for that session (excluding the grader in question). To
 175 assess intra-grader agreement across sessions, the Spearman correlation was performed between
 176 pairs of image lists ordered by a given grader at the different sessions (i.e., session 1 vs. 2, 2 vs. 3, and
 177 1 vs. 3). The average Spearman correlation coefficient with the *consensus rank* for all graders was
 178 then computed for each session (CRS1-3) and the three ranks were then averaged across the three
 179 sessions to represent an average *consensus rank* of images across all graders and sessions (CRS).

180 In experiment 2, to assess the agreement of individual grader orderings with the *consensus rank*, for
 181 each tortuosity definition T1 – T4, the Spearman correlation between each grader's ordering and the
 182 *consensus rank* of that tortuosity definition was computed, and this value was averaged across all
 183 graders for each definition (CRT1-4). To compare the correlation between different definitions of
 184 tortuosity (i.e., no specific definition in Experiment 1 and the various definitions in Experiment 2),
 185 Spearman correlations among the *consensus ranks* CRT1-4 and CRS1-3 were computed.

186

187

188

Results

Experiment 1: Grading of generic tortuosity

All ten graders completed sorting for all three sessions, resulting in a complete data set that was compiled for analysis. The correlation of the sorted list provided by each grader with the *consensus rank* of all ten graders for a given session is given in Table 1. The range in correlation coefficients was 0.43 – 0.73, with an overall mean correlation of 0.62 across all graders and sessions. Images from the data set with the lowest and highest perceived tortuosity are given in Figure 2.

	Graders										mean
	1	2	3	4	5	6	7	8	9	10	
Session 1	0.71	0.69	0.63	0.71	0.60	0.73	0.70	0.59	0.63	0.60	0.66
Session 2	0.71	0.63	0.73	0.66	0.46	0.68	0.71	0.50	0.68	0.58	0.63
Session 3	0.62	0.54	0.65	0.62	0.45	0.64	0.59	0.52	0.53	0.43	0.56
	0.68	0.62	0.67	0.66	0.50	0.69	0.67	0.54	0.61	0.54	0.62

Table 1: Spearman rank correlation coefficients for each grader and each session. Lighter/darker shading is associated with higher/lower correlation.

Figure 2. Images with the lowest (A, B) and highest (C, D) tortuosity on average, from the 30-image data set, as graded subjectively by ten expert graders in experiment 1. All images, $400 \times 400\mu\text{m}$.

To investigate intra-grader repeatability across sessions, the correlation between each grader's orderings across the different sessions was computed (Table 2), along with the average repeatability. Repeatability across sessions ranged from 0.51 – 0.89, with an overall repeatability of 0.80 across all graders. One grader, however, had a poor repeatability (mean intra-grader correlation of 0.51).

215

Graders	Sessions			mean
	1 vs 2	2 vs 3	3 vs 1	
1	0.80	0.82	0.88	0.83
2	0.93	0.88	0.87	0.89
3	0.77	0.83	0.73	0.77
4	0.78	0.68	0.75	0.74
5	0.79	0.83	0.72	0.78
6	0.96	0.95	0.95	0.95
7	0.90	0.88	0.85	0.88
8	0.79	0.84	0.86	0.83
9	0.73	0.52	0.29	0.51
10	0.76	0.75	0.59	0.70
mean				0.80

216

217

218

219

220

221

222

223

224

225

226

227

228

229

Table 2: Spearman’s rank correlation coefficients between orderings provided by the same grader in different sessions. Lighter/darker shading is associated with higher/lower correlation.

Without a specific definition of tortuosity, several graders noted difficulties in assessing and comparing nerves with numerous small-amplitude directional changes versus nerves with few large-amplitude directional changes. Another difficulty noted was whether to consider the average tortuosity of nerves in a given image or to consider the nerve with the maximum tortuosity. These difficulties were partially evident when examining the images that had the largest average sorting error across the ten observers (Figure 3). This average sorting error has been derived as the absolute difference in a grader’s rank of an image relative to the consensus rank, averaged across all graders. Nerves in these images clearly adopted non-linear paths, often with few but large-scale changes in direction and without a clear ‘start’ or ‘end’ point.

230

231

232

233

234

235

236

237

Figure 3. Images with the greatest sorting error among observers. The root mean squared error in the rank of an image from the *consensus rank* across all graders for that image was 12, 10, 7, 7, and 7 positions for images A-E, respectively (from a list with 30 possible positions corresponding to the 30-image set). All images, 400 × 400μm.

Experiment 2: Grading by specific definitions of tortuosity

For each of the four definitions of tortuosity, T1 – T4, all graders completed grading of the image set. Agreement between each grader's sorted list and the *consensus rank* for a specific tortuosity definition was assessed by the Spearman rank correlation coefficient, and is given in Table 3.

	Graders										mean
	1	2	3	4	5	6	7	8	9	10	
T1	0.62	0.21	0.68	0.56	0.59	0.71	0.55	0.67	0.60	0.60	0.58
T2	0.70	0.31	0.70	0.70	0.65	0.70	0.64	0.69	0.61	0.00	0.57
T3	0.63	0.60	0.65	0.63	0.56	0.65	0.64	0.23	0.60	0.01	0.52
T4	0.74	0.77	0.77	0.76	0.68	0.73	0.74	0.32	0.75	0.70	0.70

Table 3: Spearman rank correlation coefficient for each grader and tortuosity definition. Lighter/darker shade is associated with higher/lower correlation.

Very low correlations (<0.35) were achieved by three graders for two tortuosity definitions each. It was deemed likely that these graders misunderstood the instructions or confused the definitions of tortuosity. Across the remaining seven graders, minimum correlation was 0.55, although this value was based on inclusion of the grading results from the three graders with low correlation in computing the *consensus rank*. To overcome this limitation, it was decided to consider the very poor correlations (from graders 2, 8, and 10) as outliers and exclude their grading results from the analysis. The resulting correlations, using only data from seven graders, are given in Table 4. In this case, the range of correlation was 0.60 – 0.84, with a mean correlation across all seven graders of 0.66 – 0.70, depending on tortuosity definition. Interestingly, for all seven graders, T2 had a higher degree of correlation with the group average than T1, and similarly T4 had a higher degree of correlation compared to T3. T2 and T4 represented tortuosity definitions considering only the most tortuous nerve in the image.

	Graders							mean
	1	3	4	5	6	7	9	
T1	0.67	0.71	0.62	0.63	0.74	0.60	0.66	0.66
T2	0.81	0.84	0.83	0.78	0.84	0.76	0.72	0.80
T3	0.78	0.81	0.77	0.67	0.81	0.76	0.74	0.76
T4	0.80	0.81	0.82	0.72	0.79	0.79	0.78	0.79

Table 4: Spearman rank correlation coefficient for each grader and tortuosity definition based on seven graders.

265 To graphically represent the degree of variability in image grading among the seven graders for the
 266 four tortuosity definitions, the error in image rank for each image and grader was plotted versus the
 267 *consensus rank*, ranked from lowest to highest average tortuosity, (i.e., image 1 had lowest average
 268 tortuosity and image 30 had the highest, for a particular tortuosity definition). These graphical
 269 representations are given in Figure 4.
 270

271
 272
 273 Figure 4. Distribution of grading error for seven observers for images ordered by increasing tortuosity,
 274 based on four different tortuosity definitions T1 – T4. For each image number on the x-axis, seven
 275 black dots indicate the error in ranking of that image by each observer for the given tortuosity
 276 definition (relative to the *consensus rank* for that particular image and definition). Dashed lines are
 277 indicated at an error level of ± 10 image positions in the ranking to facilitate comparisons between the
 278 various definitions.
 279
 280 Several features became apparent when comparing the resulting distributions. The distributions
 281 confirmed that T2 and T4 had fewer outliers and better overall agreement among graders (lower

overall error) than T1 and T3, respectively. In addition, images with the lowest tortuosity and the highest tortuosity had better inter-observer agreement than images with a moderate level of tortuosity. The definitions T2 and T4 also appeared to improve the agreement in images with a moderate to high level of tortuosity, compared to T1 and T3 definitions. To numerically assess the level of agreement across the different tortuosity definitions, the consensus ranks (i.e., average ranked list of images across all seven observers) were determined and the average correlation with the consensus rank across the seven observers was computed, for each definition (Table 5).

289

	Experiment 1		Experiment 2
CRS	0.78	CRT1	0.80
		CRT2	0.86
		CRT3	0.89
		CRT4	0.90

290

Table 5: Average Spearman rank correlation coefficients between graders and the *consensus rank* for the generic tortuosity assessment in Experiment 1 (CRS, average correlation across all three sessions) and for each specific tortuosity definition in Experiment 2 (CRT, average correlation of graders with the consensus rank, for a specific tortuosity definition).

295

Observing Table 5, it is evident that a lower error for T2 and T4 noted in Table 4 and Figure 4 correspond to a higher overall inter-observer agreement for CRT2 and CRT4 relative to CRT1 and CRT3, respectively. In other words, assessing only the nerve with the highest tortuosity in the image improved the average inter-observer agreement with the *consensus rank*. Moreover, the focused definitions of tortuosity in Experiment 2 all had better inter-observer agreement compared to the subjective tortuosity assessment in Experiment 1. The *consensus rank* images for T2 and T4 definitions are shown in Figure 5. Finally, to assess the level of overlap in the generic and specific tortuosity definitions in terms of grader perceptions, Spearman correlations between all pairs of *consensus rank* image sets CRT1 – 4 and CRS1 – 3 were performed (Table 6).

305

	CRS1	CRS2	CRS3	CRT1	CRT2	CRT3	CRT 4
CRS1	1.00	0.93	0.93	0.50	0.51	0.91	0.89
CRS2	0.93	1.00	0.94	0.54	0.56	0.86	0.85
CRS3	0.93	0.94	1.00	0.63	0.66	0.86	0.87
CRT1	0.50	0.54	0.63	1.00	0.94	0.27	0.31
CRT2	0.51	0.56	0.66	0.94	1.00	0.35	0.40
CRT3	0.91	0.86	0.86	0.27	0.35	1.00	0.97
CRT4	0.89	0.85	0.87	0.31	0.40	0.97	1.00

306

307 Table 6. Spearman rank correlation between pairs of consensus ranks in Experiments 1 and 2. CRS,
 308 average correlation of graders with consensus rank for a specific session. CRT, average correlation of
 309 graders with consensus rank for a specific tortuosity definition.
 310
 311 The *consensus ranks* CRS1 – 3 were highly correlated with each other (correlation coefficient 0.93 –
 312 0.94), indicating very good reproducibility of generic tortuosity assessment between sessions. The
 313 correlation between generic assessment CRS1 – 3 and CRT1 – 2 (short range tortuosity) was relatively
 314 weak (0.50 – 0.66), whereas CRS1 – 3 correlated strongly (0.86 – 0.91) with CRT3 – 4 (long range
 315 tortuosity). For the specific definitions, CRT1 – 2 were highly correlated with each other (0.94) and
 316 similarly for CRT3 – 4 (0.97). Interestingly, however, correlation between perceptions of short and
 317 long range tortuosity were very weak (0.27 – 0.40).
 318

T2 consensus rank

T4 consensus rank

319
 320
 321 Figure 5. Images sorted by the T2 *consensus rank*, upper panels (increasing tortuosity from top left to
 322 bottom right), and images sorted by the T4 *consensus rank*, lower panels (increasing tortuosity from
 323 top left to bottom right). The positions of the images are substantially different depending on the
 324 tortuosity definition used. All images, 400 × 400µm.

Discussion

Increased tortuosity in peripheral nerves such as those of the corneal subbasal nerve plexus, may reflect pathophysiologic processes in the body that may otherwise be difficult to assess. The cornea in particular presents an ideal location for non-invasive, two-dimensional imaging of a nerve plexus within which morphologic parameters such as tortuosity can be assessed. Other techniques of peripheral nerve examination such as skin biopsies²² do not present a two-dimensional view and are therefore less amenable for investigation of long-range tortuous nerve paths. Nevertheless, despite the ease of imaging the corneal subbasal nerve plexus, standard definitions of tortuosity are lacking.

Several distinct methods to quantify subbasal nerve tortuosity have been described.^{1,4,8,15} One widely adopted method, first published by Oliveira-Soto and Efron,¹ involves a subjective, semi-quantitative grading of tortuosity on a grading scale from 0 (no tortuosity) to 4 (maximum tortuosity), exemplified by representative images containing relatively few nerves. Using this method, where the observer considers both frequency and amplitude of directional changes, a normal, healthy cornea had subbasal nerves with an average tortuosity grade of 1.2. Although the scale is simple to implement, differences in interpretation have led to variability in tortuosity values for even healthy subject groups (for example, see the comparison in Ref. 2). Although the scale has been widely used, no information is available concerning agreement between expert observers, and the scale itself was developed based only on subbasal nerves in healthy corneas.

In another common method, first described by Kallinikos et al.⁸, a numerical tortuosity coefficient was computed based on the number and frequency of directional changes derived from first and second derivatives of nerve paths. Whereas the strengths of this method are its precise mathematical definition, reproducibility, and objectivity, it is nevertheless difficult for an observer to intuitively relate the numeric coefficient to the level of tortuosity in a given image due to the use of different scaling factors. Moreover, this nerve fiber tortuosity coefficient has been shown to be highly variable in patients with diabetic neuropathy such that it has been shown to be increased, decreased and indeed unchanged compared to control subjects.^{10,12,20}

Scarpa et al.⁴ developed an alternative mathematical description of tortuosity by taking into account the number of changes in the curvature of a nerve combined with the amplitude of these changes.

354 The measure also included an empirical weighting function to boost the average tortuosity of nerves
355 in an image with one or a few very tortuous nerves. The strength of this technique is the combination
356 of an objective mathematical description with perceptions of a human observer. A weakness,
357 however, is that algorithm parameters were empirically derived based on one image set analyzed by a
358 single observer. Moreover, no information about the perception of multiple observers or inter-
359 observer variation was given.

360 Recently, another completely automated algorithm for tortuosity quantification was described by
361 Ziegler et al.¹⁵ in a population of healthy subjects and patients newly diagnosed with diabetes. The
362 measure, called CNFTo, quantified 'total absolute nerve fiber curvature'.¹⁵ Although its precise
363 definition was unclear, it did not differ between diabetic patients and control subjects, and the
364 numeric value was difficult to interpret with respect to nerve morphology.

365 Likely due to differences in interpretation and implementation of the above tortuosity measures,
366 quantifying subbasal nerve tortuosity in different pathologies has led to conflicting results. For
367 example, using the same method of Oliveira-Soto and Efron¹, significantly increased subbasal nerve
368 tortuosity was reported in both Sjögren's and non-Sjögren associated dry eye in one study,⁵ while
369 another study showed no difference in tortuosity between healthy and dry eye subjects.⁶ Sampling
370 limitations and differences in study populations, however, cannot be ruled out. Likewise, a recent
371 meta-analysis of corneal nerve assessment in patients with diabetic peripheral neuropathy concluded
372 there was no difference in nerve tortuosity between healthy or neuropathy patients or in diabetics
373 with or without peripheral neuropathy.²³ The meta-analysis also highlighted conflicting results in the
374 examined studies and a lack of standardization.

375 In order to improve the reproducibility of nerve tortuosity as a parameter describing subbasal nerves,
376 the present study sought to compare the clinical perception of tortuosity among a group of expert
377 graders. Without a specific tortuosity definition, generic grading of tortuosity was repeatable by a
378 given grader across 3 sessions (average correlation coefficient of 0.80) but the correlation among
379 different graders was low to moderate (average 0.62). The grading exercise resulted in identification
380 of two subtypes of tortuosity, describing either short-range, small amplitude or long-range, large
381 amplitude directional changes in the nerves. These specific definitions improved the inter-grader

382 agreement compared to generic grading (from 0.62 to 0.79). The short-range and long-range
383 tortuosity measures were poorly correlated with each other, suggesting that these are two distinct
384 and independent measures of tortuosity that are clearly distinguishable by expert graders. The
385 average correlation of a given grader with the *consensus rank* for the specific tortuosity definitions
386 rose to 0.86 – 0.90, when only the nerve with the maximum tortuosity in an image was considered.
387 The exclusion of additional information from other nerves in the image did not appear to affect the
388 inter-grader agreement, since the correlation between using all or only the maximum tortuosity nerve
389 was very high (0.94 – 0.97). This method therefore provides a robust means to assess a specific type
390 of tortuosity with high reproducibility and ease and speed of implementation (since only the most
391 tortuous nerve per image is assessed). This reproducibility is essential for comparing the results of
392 studies conducted in different centers. Moreover, a reproducible measure of tortuosity could also aid
393 in the possible normalization of subbasal nerve density measurements.²⁴ It is also interesting to note
394 that the specific tortuosity definitions in this study are largely independent of prior tortuosity
395 parameters. For example, the correlation between the consensus rank for T2 (CRT2) and the
396 tortuosity coefficient reported by Kallinikos et al.⁸ is 0.24.

397 The set of 30 “*consensus rank*”-ordered subbasal nerve images based on short and long-range
398 tortuosity definitions is shown in Figure 5, and can be downloaded at <http://bioimlab.dei.unipd.it> . It
399 is envisioned that these ordered image sets, representing a wide selection of tortuous nerves from
400 various pathologies and firmly based on a consensus between expert graders, will be used by
401 researchers to develop algorithms to quantify short and long-range tortuosity. This in turn can lead to
402 objective, reproducible and expert-validated automated means to assess the two distinct types of
403 nerve tortuosity in future studies.

404 While this study focused on perception and definitions of tortuosity, reproducibility in tortuosity
405 analysis can be subject to sampling bias regardless of the definition used. Further efforts are required
406 to develop a standardized sampling strategy taking into consideration the sampling area, number of
407 distinct images to use, averaging, location of images, whorl region, etc.

408 Although it remains to be seen how short-range and long-range tortuosity parameters are related to
409 subbasal nerves in healthy and pathologic corneas, it is hypothesized that one or both of these

410 distinct parameters may better describe the morphology of the nerves (and may better correlate to
411 disease) than the single tortuosity parameters used today.

412

413

References

1. Oliveira-Soto L, Efron N. Morphology of corneal nerves using confocal microscopy. *Cornea* 2001;20:374-84.
2. Patel DV, McGhee CNJ. In vivo confocal microscopy of human corneal nerves in health, in ocular and systemic disease, and following corneal surgery: a review. *Br J Ophthalmol* 2009;93:853-60.
3. Patel DV, Tavakoli M, Craig JP, Efron N, McGhee CNJ. Corneal sensitivity and slit scanning in vivo confocal microscopy of the subbasal nerve plexus of the normal central and peripheral human cornea. *Cornea* 2009;28:735-40.
4. Scarpa F, Zheng X, Ohashi Y, Ruggeri A. Automatic evaluation of corneal nerve tortuosity in images from in vivo confocal microscopy. *Invest Ophthalmol Vis Sci* 2011;52:6404-6408.
5. Benítez del Castillo JM, Wasfy MA, Fernandez C, Garcia-Sanchez J. An in vivo confocal masked study on corneal epithelium and subbasal nerves in patients with dry eye. *Invest Ophthalmol Vis Sci* 2004;45:3030-5.
6. Hoşal BM, Ornek N, Zilelioğlu G, Elhan AH. Morphology of corneal nerves and corneal sensation in dry eye: a preliminary study. *Eye* 2005;19:1276-9.
7. Labbé A, Liang Q, Wang Z, Zhang Y, Xu L, Baudouin C, Sun X. Corneal nerve structure and function in patients with non-sjogren dry eye: clinical correlations. *Invest Ophthalmol Vis Sci* 2013;54:5144-50.
8. Kallinikos P, Berhanu M, O'Donnell C, Boulton AJ, Efron N, Malik RA. Corneal nerve tortuosity in diabetic patients with neuropathy. *Invest Ophthalmol Vis Sci* 2004;45:418-22.
9. De Cilla S, Ranno S, Carini E, Fogagnolo P, Ceresara G, Orzalesi N, Rossetti LM. Corneal subbasal nerves changes in patients with diabetic retinopathy: an in vivo confocal study. *Invest Ophthalmol Vis Sci* 2009;50:5155-8.
10. Mehra S, Tavakoli M, Kallinikos PA, Efron N, Boulton AJ, Augustine T, Malik RA. Corneal confocal microscopy detects early nerve regeneration after pancreas transplantation in patients with type 1 diabetes. *Diabetes Care* 2007;30:2608-12.
11. Pritchard N, Edwards K, Shahidi AM, Sampson GP, Russell AW, Malik RA, Efron N. Corneal markers of diabetic neuropathy. *Ocul Surf* 2011;9:17-28.
12. Edwards K, Pritchard N, Vagenas D, Russell A, Malik RA, Efron N. Utility of corneal confocal microscopy for assessing mild diabetic neuropathy: baseline findings of the LANDMark study. *Clin Exp Optom* 2012;95:348-54.
13. Nitoda E, Kallinikos P, Pallikaris A, Moschandrea J, Amoiridis G, Ganotakis ES, Tsilimbaris M. Correlation of diabetic retinopathy and corneal neuropathy using confocal microscopy. *Curr Eye Res* 2012;37:898-906.
14. Maddaloni E, Sabatino F, Del Toro R, Crugliano S, Grande S, Lauria Pantano A, Maurizi AR, Palermo A, Bonini S, Pozzilli P, Manfrini S. In vivo corneal confocal microscopy as a novel non-invasive tool to investigate cardiac autonomic neuropathy in Type 1 diabetes. *Diabet Med* 2015;32:262-6.
15. Ziegler D, Papanas N, Zhivov A, Allgeier S, Winter K, Ziegler I, Brüggemann J, Strom A, Peschel S, Köhler B, Stachs O, Guthoff RF, Roden M; German Diabetes Study (GDS) Group. Early detection of nerve fiber loss by corneal confocal microscopy and skin biopsy in recently diagnosed type 2 diabetes. *Diabetes* 2014;63:2454-63.
16. Villani E, Galimberti D, Viola F, Mapelli C, Del Papa N, Ratiglia R. Corneal involvement in rheumatoid arthritis: an in vivo confocal study. *Invest Ophthalmol Vis Sci* 2008;49:560-4.

- 454 17. Schneider C, Bucher F, Cursiefen C, Fink GR, Heindl LM, Lehmann HC. Corneal confocal microscopy detects
455 small fiber damage in chronic inflammatory demyelinating polyneuropathy (CIDP). *J Peripher Nerv Syst*
456 2014;19:322-7.
- 457 18. Sturniolo GC, Lazzarini D, Bartolo O, Berton M, Leonardi A, Fregona IA, Parrozzani R, Midena E. Small fiber
458 peripheral neuropathy in Wilson disease: an in vivo documentation by corneal confocal microscopy. *Invest*
459 *Ophthalmol Vis Sci* 2015;56:1390-5.
- 460 19. Ferrari G, Grisan E, Scarpa F, Fazio R, Comola M, Quattrini A, Comi G, Rama P, Riva N. Corneal confocal
461 microscopy reveals trigeminal small sensory fiber neuropathy in amyotrophic lateral sclerosis. *Front Aging*
462 *Neurosci* 2014;6:278.
- 463 20. Hertz P, Bril V, Orszag A, Ahmed A, Ng E, Nwe P, Ngo M, Perkins BA. Reproducibility of in vivo corneal
464 confocal microscopy as a novel screening test for early diabetic sensorimotor polyneuropathy. *Diabet Med*
465 2011;28:1253-60.
- 466 21. Patel DV, McGhee CN. Quantitative analysis of in vivo confocal microscopy images: a review. *Surv*
467 *Ophthalmol* 2013;58:466-75.
- 468 22. Lauria G, Devigili G. Skin biopsy as a diagnostic tool in peripheral neuropathy. *Nat Clin Pract Neurol*
469 2007;3:546-57.
- 470 23. Jiang MS, Yuan Y, Gu ZX, Zhuang SL. Corneal confocal microscopy for assessment of diabetic peripheral
471 neuropathy: a meta-analysis. *Br J Ophthalmol* 2015 Feb 12 [epub ahead of print].
- 472 24. Edwards K, Pritchard N, Vagenas D, Russell A, Malik RA, Efron N. Standardizing corneal nerve fiber length
473 for nerve tortuosity increases its association with measures of diabetic neuropathy. *Diabet Med*
474 2014;31:1205-9.

T1, T2
short-range

T3, T4
long-range

T2 consensus rank

T4 consensus rank