Intra-and inter-observer reliability of nailfold videocapillaroscopy - A possible outcome measure for systemic sclerosis-related microangiopathy

Document Version
Accepted author manuscript

[Link to publication record in Manchester Research Explorer](#)

Citing this paper

General rights

Takedown policy

OPEN ACCESS

INTRA-AND INTER-OBSERVER RELIABILITY OF NAILFOLD VIDEOCAPILLAROSCOPY – A POSSIBLE OUTCOME MEASURE FOR SYSTEMIC SCLEROSIS-RELATED MICROANGIOPATHY

Graham Dinsdale, PhD[1], Tonia Moore, BSc[2], Neil O'Leary, PhD[3], Philip Tresadern, PhD[4],

Michael Berks, PhD[4], Christopher Roberts, PhD[3], Joanne Manning HNC[2], John Allen, PhD[5],

Marina Anderson, PhD[6], Maurizio Cutolo, MD[7], Roger Hesselstrand, MD, PhD[8], Kevin Howell,

PhD[9], Carmen Pizzorni, MD, PhD[7], Vanessa Smith, MD, PhD[10], Alberto Sulli, MD[7], Marie

Wildt[8], Christopher Taylor, PhD[4], Andrea Murray, PhD[1], Ariane L Herrick, MD[1,11]

1. Division of Musculoskeletal & Dermatological Sciences, University of Manchester, Manchester Academic Health Science Centre, Manchester, UK
2. Salford Royal Hospital NHS Foundation Trust, Salford, UK
3. Centre for Biostatistics, Division of Population Health, Health Services Research & Primary Care, University of Manchester, Manchester, UK
4. Centre for Imaging Sciences, Division of Informatics, Imaging & Data Sciences, University of Manchester, Manchester, UK
5. Microvascular Diagnostics, Northern Medical Physics and Clinical Engineering, Freeman Hospital, Newcastle upon Tyne, UK
6. Institute of Ageing and Chronic Disease, University of Liverpool, Liverpool, UK
7. Research Laboratory and Academic Division of Clinical Rheumatology, Dept. Internal Medicine, University of Genova, Italy

8. Department of Clinical Sciences, Section of Rheumatology, Lund University, Lund, Sweden

9. Institute of Immunity and Transplantation, University College London, Royal Free Campus, London, UK

10. Department of Rheumatology, Ghent University Hospital, Faculty of Internal Medicine, Ghent University, Ghent, Belgium

11. NIHR Manchester Musculoskeletal Biomedical Research Unit, Central Manchester NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK.

Corresponding author: Dr Graham Dinsdale, C214 Clinical Sciences Building, Salford Royal Hospital, Stott Lane, Salford, M6 8HD, UK. Email: graham.dinsdale@manchester.ac.uk. Telephone: (+44) 161 206 2935

Word count (excluding title page, abstract, references, figures and tables): 3,012

Competing interests: The authors declare no conflict of interests.

ABSTRACT

Objectives

Our aim was to assess the reliability of nailfold capillary assessment in terms of image evaluability, image severity grade ('normal', 'early', 'active', 'late'), capillary density, capillary (apex) width, and presence of giant capillaries, and also to gain further insight into differences in these parameters between patients with systemic sclerosis (SSc), patients with primary Raynaud's phenomenon (PRP) and healthy control subjects.

Methods

Videocapillaroscopy images (magnification 300x) were acquired from all 10 digits from 173 participants: 101 patients with SSc, 22 with PRP and 50 healthy controls. Ten capillaroscopy experts from 7 European centres evaluated the images. Custom image mark-up software allowed extraction of the following outcome measures: overall grade ('normal', 'early', 'active', 'late', 'non-specific', or 'ungradeable'), capillary density (vessels/mm), mean vessel apical width, and presence of giant capillaries.

Results

Observers analysed a median of 129 images each. Evaluability (i.e. the availability of measures) varied across outcome measures (e.g. 73.0% for density and 46.2% for overall grade in patients with SSc). Intra-observer reliability for evaluability was consistently higher than inter- (e.g. for density, intra-class correlation coefficient [ICC] was 0.71 within and 0.14 between observers). Conditional on evaluability, both intra- and inter-observer reliability were high for grade (ICC 0.93 and 0.78 respectively), density (0.91 and 0.64) and width (0.91 and 0.85).

Conclusions

Evaluability is one of the major challenges in assessing nailfold capillaries. However, when images are evaluable, the high intra- and inter-reliabilities suggest that overall image grade, capillary density and apex width have potential as outcome measures in longitudinal studies.

KEYWORDS

- Systemic sclerosis

- Nailfold videocapillaroscopy

- Reliability

- Capillaroscopic patterns

- Raynaud's phenomenon

INTRODUCTION

Nailfold capillaroscopy allows direct, non-invasive visualisation of the digital microcirculation, which is abnormal in the multisystem connective tissue disease systemic sclerosis (SSc) [1,2]: abnormalities include widened, architecturally abnormal capillaries and areas of avascularity[3,4]. A major implication of the inclusion of abnormal nailfold capillaroscopy into the 2013 American College of Rheumatology (ACR)/European League Against Rheumatism (EULAR) classification criteria for SSc [5,6], is that it now behoves all rheumatologists to have some understanding of the technique and of image interpretation (including the challenges and caveats therein). To date, the main application of nailfold capillaroscopy has been the identification and prediction of SSc-spectrum disease in patients with Raynaud's phenomenon [7-9], separating this from primary Raynaud's phenomenon (PRP) in which the nailfold capillaries are normal. Because Raynaud's phenomenon is often the first symptom of SSc (a disease with high morbidity and mortality), nailfold capillaroscopy therefore affords a window of opportunity for early diagnosis (and intervention). However, in the research arena there is increasing interest in capillaroscopy as an outcome measure in clinical studies of SSc-related microvascular disease. Advances in hardware, and in image annotation software [10-12], mean that nailfold capillary changes can be tracked over time, affording new insights into pathophysiology and potentially reliable measurement of change.

A key requirement of any method for application in either clinical practice or research is that of reliability. We need to know the level of agreement between observers when reporting those parameters currently used to assess (the degree of) abnormality: these include vessel density (number of distal row capillaries/mm of nailfold), capillary width (measured at the

capillary apex), presence of giant capillaries (homogenously enlarged capillaries with a vessel diameter of >50um, generally considered pathognomonic of a SSc-spectrum disorder), and image grading according to the 'early', 'active' and 'late' classification of Cutolo et al [10]. Importantly, it must also be recognised that the distal row of capillaries cannot always be visualised, sometimes due to the physical limitations of imaging in optically dense tissue (meaning that a significant percentage of incident light is lost via scattering) and sometimes (in patients with SSc) because the nailfold is so abnormal (and so avascular) that no capillaries are present or can be adequately seen. Whether or not a nailfold capillary image is evaluable can be subjective, and so evaluability is a further parameter to take into account when assessing reliability [13].

Against this background, our aim was to assess the reliability of nailfold capillary assessment in terms of image evaluability, image grade (early, active, late), capillary density, capillary (apex) width, and presence of giant capillaries, and also to gain further insight into differences in these parameters between patients with SSc, patients with PRP and healthy control subjects.

PATIENTS AND METHODS

Patients

A total of 173 participants were recruited into the study (101 patients with SSc, 22 with PRP and 50 healthy controls) at Salford Royal NHS Foundation Trust, a tertiary referral centre for SSc (Table 1). Patients included in the SSc group all fulfilled the 2013 American College of Rheumatology/European League Against Rheumatism criteria for SSc [5,6]. Patients included in the PRP group had no clinical features of connective tissue disease, were antinuclear antibody negative (≤ 1/100) and had no previous abnormal nailfold capillaroscopy findings. All imaging was carried out after 20 minutes acclimatisation at 23°C in a temperature and humidity-controlled laboratory. Participants were asked to refrain from caffeine and smoking for 4 hours before each study visit. All participants gave informed, written consent. The study was approved by the Greater Manchester East Research Ethics committee (reference: 11/NW/0444).

Table 1 – Participant demographics, and numbers of images acquired and evaluated

	Control (n=50)	PRP (n=22)	SSc (n=101)
Median age [range] (years)	44 [19-71]	46 [20-69]	62 [26-80]
Female [%]	31 [62%]	18 [82%]	85 [84%]
Current smoker [%]	4 [8%]	5 [23 %]	9 [9%]
Limited cutaneous subtype [%]	-	-	81 [80%]
Median time since RP* onset [range] (years)	-	9 [1-45]	15 [0-64]
Median time since 1st non-RP* symptom [range] (years)	-	-	12 [0-35]
History of IV iloprost use [%]	-	-	16 [16%]
History of debridement [%]	-	-	7 [7%]
History of amputation [%]	-	-	0 [0%]
Current vasodilator [%]	-	6 [27%]	39 [39%]
Number of images	496	209	945
Number of evaluations	863	552	1986
*RP: Raynaud's phenomenon			

Image acquisition

Panoramic nailfold videocapillaroscopy imaging was performed on all 10 digits (including thumbs) using a microscope with 300x magnification (approx. 1um/pixel resolution) (KK Technology, Honiton, UK), and green LED illumination for maximum capillary contrast. Custom capture software allowed images to be recorded as panoramic mosaics across the entire nailfold by registering and digitally stitching individual camera frames [11,12]. Capturing panoramic mosaics in this manner insures against the possibility of false assessment due to the spatial heterogeneity of the capillary architecture, as shown in Figure 1.
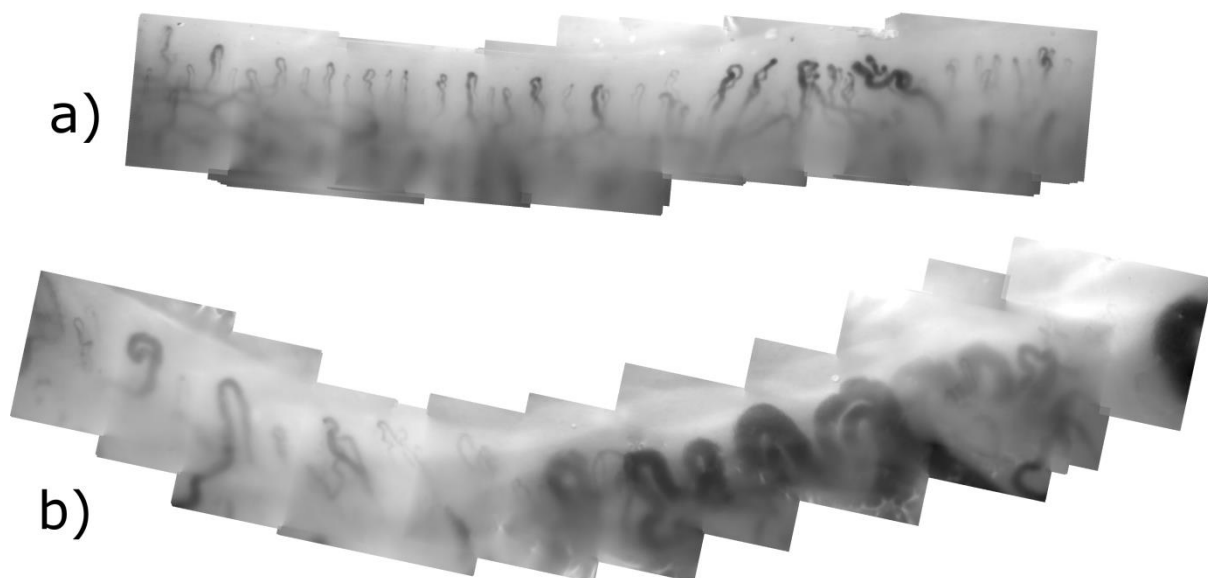
Figure 1. Example mosaic nailfold images. These demonstrate the heterogeneity of capillary architecture and why it is important to view the whole nailfold. Panel a) displays normal

capillaries on the left but some avascularity, capillary enlargement and distortion of the vessel bed on the right. The lower panel, b), shows avascularity and evidence of neoangiogenisis on the left, while on the right there are mainly greatly enlarged or giant vessels with generally regular shape and spacing. In both examples, consideration of only one half of the nailbed would give a different 'result' from the other half.

Image analysis and grading

A group of 10 capillaroscopy experts ('observers') from seven European centres assessed the images. Custom image mark-up software was written, allowing detailed, quantitative image assessment using a set of pre-determined criteria (described in detail below). The software allowed image batches to be delivered to each observer, and results to be returned securely, via the internet. In order to reduce the possibility of any training effects, observers assessed a test batch of 25 video capillaroscopy images (taken from subjects in a previous capillaroscopy study) in order to familiarise themselves with the software interface and scoring protocol. Following successful completion of the training, blinded observers (blinded to all subject characteristics including finger and 'patient' group) were randomly assigned batches (stratified by patient group, and with sufficient overlap of images between observers to allow assessment of inter-observer reliability) of around 50 images at a time (including some repeat images to allow assessment of intra-observer reliability); on completion further images were then made available to them. Two observers were vascular technicians (from two different centres) and (by prior agreement) were sent larger numbers of images than the other observers. Observers were presented with each videocapillaroscopy image in turn (Figure 2) and asked to: (1) assess overall image grade, (2) mark the location of distal capillary apices, (3) categorically grade the marked (distal) vessels in terms of size ('Normal', 'Enlarged', 'Very enlarged' ['Giant' = homogeneously enlarged, or 'irregularly enlarged']) and shape ('Normal', 'Non-specific', or 'Angiogenic'), and (4) measure the apical width of marked (distal) vessels using a click-and-drag tool. When assessing the overall image grade, observers could choose from 'Normal', 'Early', 'Active', 'Late', or 'Non-specific' (non-specific changes in vessel shape and/or size) categories. There were also two

'Ungradeable' categories for: (1) images considered to be so abnormal as to be unclassifiable/unmeasurable and (2) poor image quality. It should be noted that vessels marked as "Giant" in size, were automatically assigned a "Normal" shape.
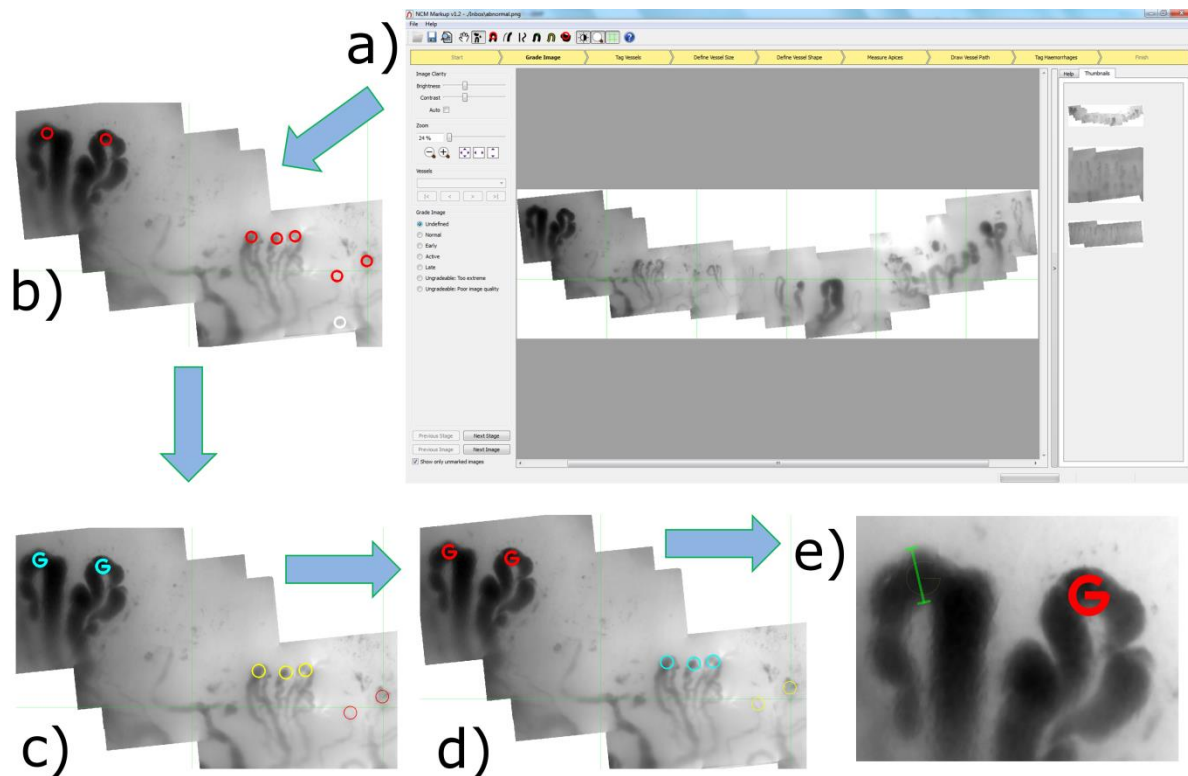
Figure 2. Step-by-step nailfold image markup process. Steps: a) the overall software interface displaying a nailfold mosaic – the first step of markup is where the observer grades the image overall; b) vessel apex identification – red circles are the distal vessels identified for further analysis; c) vessel size categorisation – each identified distal vessel is assigned a size ("Giant": light blue 'G' marker, "Irregularly enlarged": light blue circle [not demonstrated here], "Enlarged": yellow circle, "Normal": red circle); d) vessel shape categorisation – each vessel is assigned a shape ("Giant": red 'G' marker, "Normal": red circle [not shown here], "Non-specific": yellow circle, "Angiogenic": light blue circle); e) apex width measurement – apices are measured by the user clicking and dragging a green marker across the area.

Outcome measures and measure requirements

From the returned dataset, four image-level outcome measures were extracted or derived to allow further analysis: (1) the observer's overall grade (hereafter *grade*) of the image on a 4-point severity scale ranging from "normal" (assigned a value of 1), through "early"(assigned 2) and "active" (assigned 3), to "late" (assigned 4). Images graded as "non-specific", or in either of the "ungradeable categories" were not considered; (2) capillary density (hereafter *density*; vessels/mm) was calculated as the total number of distal vessels divided by the Euclidean distance between the leftmost and rightmost vessels marked in the image. Density was only calculated when 2 or more vessels where identified within an image; (3) mean apex width (hereafter *width*) was defined as the arithmetic mean of the measured apex width of all vessels measured in an image, conditional on at least one distal vessel being identified and measured; (4) presence of giant capillaries (hereafter *giants*) was defined as a binary variable, set TRUE if there was one or more vessels identified and subsequently categorically marked as having size/shape in the "Giant" category, set FALSE otherwise.

Statistical analysis

Reliability can be measured using an intra-class correlation coefficient, which expresses the between subject variance as a proportion of the total variance. The dataset was a complex structure with nailfolds nested both within subjects and observers. Numbers of images for each subject also varied. For this reason variance components were estimated using a linear mixed model with crossed random effects. All models included observer and patient variation modelled with random effects of finger nested within patient and within observer.

In each evaluation, evaluability of the four outcome measures was recorded as a binary variable indicating whether the measure requirement had been met (i.e. measure availability). A mixed-effects model was used to estimate the mean probability of an observer providing a usable value of each measure for an image. A low availability or probability indicates a poor measure.

Reliability and validity of measures can only be assessed were data is available. These analyses were therefore based on conditional models, i.e. including only images with a usable value being provided. All models included covariates for finger, gender and age as fixed-effects to account for systematic non-pathological differences. Models for reliability estimates did not include any further fixed-effects, so between-patient variance components could be estimated without reducing variance from severity of disease. Reliability estimates thus reflect disease-related variation. Including fixed effects of hand and fingers means these correlations are the average reliability stratified by finger and hand. Validity and evaluability models additionally included disease group as a fixed-effect. A Likelihood-ratio omnibus test was used to assess the evidence for differences across diagnosis groups for each outcome measure. A semi-parametric bootstrap was used to derive confidence intervals of estimates and differences between them; 1000 bootstrap samples were used for each procedure. Analyses were performed using R [14], and the package lme4 [15].

RESULTS

Numbers of images graded

The observers analysed a median (interquartile range) of 129 (117-167) images each. In total, the pool of 10 observers returned 3401 image evaluations from 1650 unique nailfold images, with 2731 evaluations (80%) having 2 or more vessels identified. Table 2 details the number of images that each observer evaluated, broken down by diagnosis category and percentage proportion of each observer's total. Although observers evaluated very different numbers of images, the division between the diagnosis categories was broadly similar for all.

Table 2 – Number and proportion of images evaluated at baseline by observer and diagnosis

Observer	Healthy Control (number [observer percentage])	PRP (number [observer percentage])	SSc (number [observer percentage])	Total	Proportion of total evaluation pool (%)
1	46 (27.2)	32 (18.9)	91 (53.8)	169	5.0
2	18 (18.4)	20 (20.4)	60 (61.2)	98	2.9
3	20 (16.7)	20 (16.7)	80 (66.7)	120	3.5
4	20 (17.5)	20 (17.5)	74 (64.9)	114	3.4
5	26 (21.8)	20 (16.8)	73 (61.3)	119	3.5
6	36 (26.3)	37 (27.0)	64 (46.7)	137	4.0
7	228 (31.3)	92 (12.6)	408 (56.0)	728	21.4
8	20 (17.2)	21 (18.1)	75 (64.7)	116	3.4
9	416 (25.4)	253 (15.4)	972 (59.2)	1641	48.3
10	33 (20.8)	37 (23.3)	89 (56)	159	4.7
Total image evaluations (mean observer percentage)	863 (22.3)	552 (18.7)	1986 (59.0)	3401	100.0

Of the 1650 unique, imaged nailfolds evaluated by at least one observer, 358 (22%) images had a repeat evaluation by the same observer and 904 (57%) were evaluated by at least two observers, with a total of 951 (59%) images contributing directly either to intra-observer or inter-observer reliability parameter estimates for evaluability. Due to time constraints, it was not possible to have all 1650 images assessed by 2 or more observers.

Evaluability (measure availability)

Table 3(i) shows the adjusted measure availability (as a percentage), along with 95% confidence intervals, for each of the four outcome measures, subdivided by diagnosis. The p-values in Table 3 relate to the omnibus test that all three groups are different, rather than any particular pair-wise test between the three groups. Availability of measures varied widely between observers. For outcome measures 2 and 4 (density and giant vessels), the range was from 41.0% to 98.0% across the three diagnosis groups (P=0.06). For outcome measure 3 (width), it was from 38.9% to 96.5% (P=0.12) and for measure 1 (grade), it was from 18.5% to 78.1% (P=0.81). One observer was consistently and appreciably lower in these rates than others, excluding this observer gave a minimum rate of availability for the above figures of 34.3%, 52.0%, 51.7%, and 52.0%.

For the 3401 image evaluations producing an overall image grade, "Non-specific" was the most frequent category: 34.6%, 28.3% and 38.9% for healthy controls, PRP and patients with SSc respectively, with an "Ungradeable quality" given for 18.5%, 20.0% and 15.0% of image evaluations for the respective diagnosis groups.

Table 3 – (i) Adjusted (mean) probability for availability of each outcome measure, (ii) Adjusted mean value by participant disease category. Fixed-effects estimates are from mixed-effects linear regression on each image (conditional on each measure being measurable - (ii) only)

	Healthy Control	95% CI	PRP	95% CI	SSc	95% CI	P (omnibus)
(i) Measure availability (%)							
1. Overall image grade*	47.1	34.3, 59.9	44.3	32.0 56.6	46.2	32.2, 60.2	0.81
2. Distal vessel density	77.9	64.0, 91.9	82.0	67.5, 96.5	73.0	59.3, 86.8	0.06
3. Mean distal vessel width	76.0	61.8, 90.1	79.6	64.5, 94.6	70.2	56.4, 84.0	0.12
4. Presence or absence of giant capillary	77.9	64.0, 91.9	82.0	67.5, 96.5	73.0	59.3, 86.8	0.06
(ii) Outcome measure (conditional on availability)							
1. Overall image grade*	1.15	0.90, 1.41	1.27	0.93, 1.61	2.88	2.65, 3.10	<0.001
2. Distal vessel density (vessel/mm)	10.1	9.2, 10.9	9.7	8.7, 10.7	6.3	5.5, 7.1	<0.001
3. Mean distal vessel width (μm)	11.6	9.2, 14.3	13.4	10.1, 16.8	21.2	19.1, 23.4	<0.001
4. Presence of giant capillary (%)	4.8	-4.8, 14.4	6.0	-5.0, 17.0	24.2	15.1, 33.2	<0.001

*Grade severity: 1 (normal), 2 (early), 3 (active), 4 (late)

(All CI's, normal approximations, p-values from likelihood ratio test)

Estimates are expectations at the mean age of 53.9 years, averaged across left/right hands and digits (assuming equal proportions).

Reliability

Table 4 shows the intra- and inter-observer reliability of both measure availability and measurement, for all four outcome measures.

For measure availability, intra-observer reliability was consistently much higher than inter-observer reliability. For reliability of the four outcome measures (conditional on evaluability), again intra-observer was higher than inter-observer reliability, but only marginally so for mean vessel width and for overall image grade.

Table 4 – Intra-class correlation coefficients for measure availability and vessel mark-up

	Intra-observer	Inter-observer
Measure availability		
Overall image grade	0.61	0.22
Distal vessel density	0.71	0.14
Mean distal vessel width	0.78	0.26
Presence or absence of giant capillary	0.71	0.14
Measure (conditional on availability)		
Overall image grade	0.93	0.78
Distal vessel density	0.91	0.64
Mean distal vessel width	0.91	0.85
Presence of giant capillary	0.71	0.41

Outcome measures across clinical groups

Table 3(ii) shows the adjusted mean values for the four outcome measures, again subdivided by diagnosis. These measures are conditional on their availability. After adjustment the overall image grade (on a 1=normal to 4=late scale) was 0.11 units higher in patients with PRP and 1.72 units higher in patients with SSc, compared to controls (P<0.001). Mean vessel density was lower in both PRP and SSc groups by 0.4 vessels/mm and 3.8 vessels/mm, respectively, compared to healthy controls (P<0.001). The adjusted mean vessel width was 1.7 µm higher in patients with PRP, and 9.5 µm higher in patients with SSc compared with healthy controls (P<0.001).

In images judged to be evaluable, the presence of giant capillaries was noted in 4.8%, 6.0%, and 24.2% of image observations for healthy controls, and patients with PRP and SSc, respectively (P<0.001).

There was a clear association between overall image grade and diagnosis. When image grade was provided on the severity scale (normal, early, active or late), it was likely to be more severe in patients with SSc (16.9% evaluated as normal) compared to healthy controls (89.2% normal) and patients with PRP (79.4% normal). Severity was more frequently observed as active (35.1%) or late (33.2%) in patients with SSc than as normal (16.9%) or early (14.8%).

DISCUSSION

Our study involved seven different European centres supplying ten observers who examined 1650 images covering a broad spectrum of normality and abnormality. The study highlights the challenges and complexities of nailfold capillaroscopy image assessment. A key finding was the high intra-observer reliability (conditional on evaluability) of overall image grade, capillary density and apex width. Inter-observer reliability was also good, although lower, for these measures, suggesting that they have potential as outcome measures in both single-centre (i.e. single-observer) and multi-centre studies. However, our study has also highlighted the real-world situation that some capillaroscopy images are judged as unevaluable, with some disagreement amongst observers about evaluability. The reliability findings above are conditional on capillaries being sufficiently well-visualised to allow assessment. It should be noted, however, that observers assessed density in 80.3% of evaluations and apex width in 76.1% of evaluations, suggestive of observer confidence when using these quantitative measures. It is likely that evaluability is to some extent dependent upon the task given to the observer. In a recent study from our department, 84% of images could be classified, although observers were only asked to grade capillaries as 'normal', 'mildly', 'definitely' or 'grossly' abnormal [13]. In the present study, the task was more complex.

The study also confirmed the age and sex-adjusted differences in capillary morphology between patients with SSc, patients with PRP, and healthy controls, although this was not a primary objective.

Various previous studies report reliability of high magnification nailfold capillaroscopy, highlighting the implied need for standardisation [12,13,16-24]. Some included only two

21

observers [12,16,17,20,21]. A small number of these studies included assessment of reliability of capillary density and/or apical width. Hudson et al reported high intra- and inter-observer (two observers) reliability of density using a widefield microscope (50x magnification) [25]. We previously reported high inter- and intra-observer reliability of quantitative parameters in a six-observer/three-centre study examining panoramic nailfold mosaics: ICCs (intra- and inter-observer, for evaluable images) for capillary density were 0.92 and 0.71 respectively and for apical width 0.97 and 0.94 respectively [18]. Sekiyama et al reported ICCs for intra- and inter-observer reliability (two observers) for number of capillaries per mm of 0.991 and 0.952 respectively for videocapillaroscopy, and 0.993 and 0.954 respectively for widefield capillaroscopy (10-25x magnification) [21]. Each observer acquired new images (from all eight fingers), and so, unlike our current study, reliability incorporated both image acquisition and analysis [21]. The current study had the advantage of being based on a large panel of observers, with analysis including finger, patient gender and age as covariates, thereby removing differences between images and subjects that relate to these characteristics. A clinical trial would likely also control for these characteristics, so our estimates of reliability may be more realistic for clinical trials. Together, the current study, and those described, support capillary density and apical width as reliable, quantifiable outcome measures.

Smith et al studied reliability of assessment of the 'early', 'active', 'late' grading (patient-level data from eight fingers) [17]. A high proportion of images (almost half by one observer) were rated 'non-SSc-specific', meaning images were either normal or 'non-specific'. Given that all patients in the study had SSc it seems likely that most were therefore 'non-specific', consistent with the current study findings. In this current study, again a high

proportion of images was graded 'non-specific', meaning that although the image was not normal, it did not fit neatly into the ordered grading scale of 'early', 'active' or 'late' and was therefore excluded from the analysis, which used a numerical scoring system based on image grading (1 for 'normal' through to 4 for 'late'). For this reason, quantitative assessment for example with capillary density or width has advantages over a qualitative scoring system.

A very recent study from Smith et al reported that when observers were asked to assess a single capillary ('normal', 'abnormal' or 'not evaluable') an unspecified proportion of the images were ranked 'not evaluable' [23]. This indicates the complexities of capillaroscopic analysis, and how challenging interpretation can be at the individual capillary level as well as at both the 'patient' level (averaging across fingers) [17,21] and image/'finger' level (as in the current study). A further challenge is the heterogeneity of 'normal' appearances. Although it was not within the remit of this study to examine the variation in nailfold capillary appearances across healthy control subjects, it is worth noting, for example, that even in healthy controls, raters subjectively marked a small proportion of capillaries as 'giants'. It is worth highlighting the importance of understanding reliability for outcome measures derived from nailfold capillaroscopy. For clinical practice, a very high level of inter-observer reliability is needed for a measure to be used for diagnosis. Such high levels are often difficult to achieve and so measures are interpreted alongside other information when assessing the patient. However, in clinical trials, high levels of inter-observer reliability are not essential. For example if intra-observer reliability is substantially higher than inter-observer reliability, a single observer could be used although this may lack generalisability.

Alternatively, the assessment process could be stratified by observer so that observer is not confounded with treatment (i.e. similar proportions from each treatment arm are assigned to each observer). Another option is to use a panel of observers, with the average rating used as the outcome. This improves both generalisability and reliability. The reliability of a panel of independent observers can be determined from the inter-observer reliability using the Spearman-Brown formula [26]. The inter-observer reliability for vessel density in the current study was 0.64. If the outcome was e.g. vessel density averaged across four independent observers, reliability would be 0.88 by this formula. An increased reliability reduces the required sample sizes for clinical trials. Using an average across four observers would then reduce the required sample size to 0.64/0.88; a 28% reduction in participant numbers. See Fleiss (1986) for further discussion of these matters [27].

Our study had limitations. Its aim was to assess reliability of image interpretation only (not including image acquisition). Reliability of outcome measures could only be estimated using data from images judged as evaluable; reflecting the reality of the clinical setting. Observers could choose which segments of nailfold to consider, although they were expected to analyse as much of the panoramic nailfold as possible. A key challenge of capillaroscopic assessment is the heterogeneity of capillary morphology even within one nailfold (Figure 1), highlighting that the same section of nailfold should be examined each time [20]. Additionally, observers received no specific training (shown to improve reliability of capillaroscopic assessment [19]) although all had extensive clinical and research experience with capillaroscopy. Finally, although numbers included in the PRP group were small, the main purpose of the study was to assess reliability of assessors in grading images. The PRP

group was included primarily to ensure that the full spectrum of nailfold appearance was presented to the assessors, ranging from normal to extremely abnormal.

In summary, nailfold capillaroscopic analysis is challenging, in part due to the difficulty in obtaining good quality images from all nailfolds and the variability between observers in determining what constitutes an evaluable nailfold. However, capillary density and apical width, parameters likely to be helpful in quantifying changes over time, were reliable measures in evaluable nailbeds. Quantitative capillaroscopic parameters were recently confirmed of value in prediction of disease progression.[28-29] We are therefore optimistic that, through a combination of standardised training and increased use of automated analysis [30], the non-invasive technique of videocapillaroscopy will ultimately deliver as an outcome measure of SSc-related microangiopathy.

REFERENCES

1. Campbell PM, LeRoy EC. Pathogenesis of systemic sclerosis: a vascular hypothesis. Semin Arthritis Rheum 1975: 4: 351-68.

2. Herrick AL. Vascular function in systemic sclerosis. Curr Opin Rheumatol 2000; 12: 527-33.

3. Maricq HR, LeRoy EC. Patterns of finger capillary abnormalities in connective tissue disease by 'wide-field' microscopy. Arthritis Rheum 1973; 16: 619-28.

4. Cutolo M, Grassi W, Matucci Cerinic M. Raynaud's phenomenon and the role of capillaroscopy. Arthritis Rheum 2003; 11: 3023-30.

5. Van den Hoogen F, Khanna D, Fransen J, et al. 2013 Classification criteria for systemic sclerosis. Arthritis Rheum 2013; 65: 2737-47.

6. Van den Hoogen F, Khanna D, Fransen J, et al. 2013 Classification criteria for systemic sclerosis. Ann Rheum Dis 2013; 72: 1747-55.

7. Koenig M, Joyal F, Fritzler MJ, et al. Autoantibodies and microvascular damage are independent predictive factors for the progression of Raynaud's phenomenon to systemic sclerosis: a twenty-year prospective study of 586 patients, with validation of proposed criteria for early systemic sclerosis. Arthritis Rheum. 2008;58:3902-12.

8. Ingegnoli F, Boracchi P, Gualtierotti R, et al. Improving outcome prediction of systemic sclerosis from isolated Raynaud's phenomenon: role of autoantibodies and nail-fold capillaroscopy. Rheumatol. 2010; 49, 797–805.

9. Herrick AL, Cutolo M. Clinical implications from capillaroscopic analysis in patients with Raynaud's phenomenon and systemic sclerosis. Arthritis Rheum 2010; 62: 2595-604.

10. Cutolo M, Sulli A, Smith V. How to perform and interpret capillaroscopy. Best Practice Res Clin Rheumatol 2013; 27: 237-48.

11. Allen PD, Taylor CJ, Herrick AL, et al. Image analysis of nailfold capillary patterns from video sequences.   Proceedings of Medical Image Computing and Computer-Assisted Intervention, Springer, 1999; 698-705.

12. Anderson ME, Allen PD, Moore T, et al. Computerized nailfold video capillaroscopy - a new tool for assessment of Raynaud's phenomenon. J Rheumatology 2005; 32: 841-8.

13. Hughes M, Moore T, O'Leary N, et al. A study comparing videocapillaroscopy and dermoscopy in the assessment of nailfold capillaries in patients with systemic sclerosis-spectrum disorders. Rheumatology 2015; 54: 1435-42.

14. R Core Team. R: A Language and Environment for Statistical Computing [computer program]. Vienna, Austria: Foundation for Statistical Computing, 2013. http://www.r-project.org/.

15. Bates D, Mächler M, Bolker B, et al. Fitting linear mixed-effects models using lme4. J Stat Softw 2015; 67(1), 1-48. doi:10.18637/jss.v067.i01.

16. Ingegnoli F, Gualtierotti R, Lubatti C, Zahalkova L, Meani L, Boracchi B, et al. Feasibility of different capillaroscopic measures for identifying nailfold microvascular alterations. Semin Arthritis Rheum 2009; 38: 289-95.

17. Smith V, Pizzorni C, De Keyser F, et al. Reliability of the qualitative and semiquantitiave nailfold videocapillaroscopy assessment in a systemic sclerosis cohort: a two-centre study. Ann Rheum Dis 2010; 69: 1092-6.

18. Hofstee HMA, Serne EH, Roberts C, et al. A multicentre study on the reliability of qualitative and quantitative nail-fold videocapillaroscopy assessment. Rheumatol 2012; 51 : 749-55.

19. Gutierrez M, Bertolazzi C, Tardella M, et al. Interreader reliability in assessment of nailfold capillary abnormalities by beginners: pilot study of an intensive videocapillaroscopy training programme. J Rheumatol 2012; 39: 1248-55.

20. Murray AK, Vail A, Moore TL, et al. The influence of measurement location on reliability of quantitative videocapillaroscopy in patients with SSc. Rheumatol 2012; 51 : 1323-30.

21. Sekiyama JY, Camargo CZ, Eduardo L, et al. Reliability of widefield nailfold capillaroscopy and videocapillaroscopy in the assessment of patients with Raynaud's phenomenon. Arthritis Care Res 2013; 65: 1853-61.

22. Overbury R, Murtaugh MA, Fisher A, et al. Primary care assessment of capillary abnormalities in patients with Raynaud's phenomenon. Clin Rheumatol 2015; 34: 2135-40.

23. Smith V, Beeckman S, Herrick AL, et al. 1 An EULAR study group pilot study on reliability of simple capillaroscopic definitions to describe capillary morphology in rheumatic diseases. Rheumatol 2016; 55: 883-90.

24. Cheng C, Lee CW, Deskalakis C. A reproducible computerized method for quantitation of capillary density using nailfold capillaroscopy. J Vis Exp 2015; 105: doi: 10.3791/53088.

25. Hudson M, Masetto A, Steele R, et al. Reliability of widefield capillary microscopy to measure nailfold capillary density in systemic sclerosis. Clin Exp Rheumatol 2010; 28 (5 Suppl 62): S36-S41.

26. Spearman, CC C. Correlation calculated from faulty data. *Br J Psychol* 1910; 3: 271–95.

27. Fleiss, J L. Design and Analysis of Clinical Experiments. New York: John Wiley & Sons 1986.

28. Trombetta AC, Smith V, Pizzorni C et al. Quantitative alterations of capillary diameter have a predictive value for development of the capillaroscopic systemic sclerosis pattern. J Rheumatol. 2016; 43(3):599-606.

29. Cutolo M, Herrick AL, Distler O et al. Nailfold videocapillaroscopic and other clinical risk factors for digital ulcers in systemic sclerosis: a multicenter, prospective cohort study. Arthritis Rheumatol. 2016 Apr 25. doi: 10.1002/art.39718

30. Berks M, Tresadern P, Dinsdale G, et al. An automated system for detecting and measuring nailfold capillaries. Medical Image Computing & Computer-Assisted Intervention: MICCAI. 2014; 17: 658-65.