

**A study examining the reliability of digital ulcer definitions as proposed by the UK Scleroderma Study Group: challenges and insights for future clinical trial design**

Concise report

Michael Hughes<sup>1</sup>, Andrew Tracey<sup>2</sup>, Monica Bhushan<sup>3</sup>, Kuntal Chakravarty<sup>4</sup>, Christopher P Denton<sup>5</sup>, Serena Guiducci<sup>6</sup>, Shirish Dubey<sup>7</sup>, Voon Ong<sup>5</sup>, Louise Parker<sup>5</sup>, John Pauling<sup>8</sup>, Athiveer Prabu<sup>9</sup>, Janet Ravenscroft<sup>10</sup>, Christine Rogers<sup>11</sup>, Christopher Roberts<sup>12</sup>, Ariane L Herrick<sup>1,13</sup>

Corresponding author:

Dr Michael Hughes.

Centre for Musculoskeletal Research, The University of Manchester, Salford Royal NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK.

Michael.hughes-6@postgrad.manchester.ac.uk

Telephone: +44 (0)161 206 2460

Abstract = 237/250

Word count = 1962/2000

1 Figure & 1 Table (maximum allowed)

## **Abstract**

**Objectives:** The reliability of clinician grading of systemic sclerosis (SSc)-related digital ulcers (DUs) has been reported to be poor to moderate at best, which is a major concern in clinical trial design. The aim of this study was to examine the reliability of new proposed UK Scleroderma Study Group DU definitions amongst UK clinicians with an interest in SSc.

**Methods:** Raters graded (through a custom-built interface) 90 images (80 unique and 10 repeat) of a range of digital lesions collected from patients with SSc. Lesions were graded on an ordinal scale of severity: 'no ulcer', 'healed ulcer', or 'DU'.

**Results:** Twenty-three clinicians: 18 rheumatologists, 3 dermatologists, one hand surgeon and one specialist rheumatology nurse, completed the study. A total of 2070 (1840 unique + 230 repeat) image gradings were obtained. For intra-rater reliability, the overall weighted kappa coefficient was high (0.71) and was moderate (0.55) when averaged across raters. Overall inter-rater reliability was poor (0.15).

**Conclusion:** Although our proposed DU definitions had high intra-rater reliability, the overall inter-rater reliability was poor. Our study further highlights the challenges of DU assessment by clinicians with an interest in SSc, and also provides a number of useful insights for the design of future clinical trials. Further research is warranted to improve the reliability of DU definition/rating as an outcome measure in clinical trials, including examining the role for objective measurement techniques, and the development of DU patient reported outcome measures.

### **Key messages**

- Reliable methods of DU assessment are needed to facilitate future clinical trials.
- Our proposed UKSSG DU definitions had good intra-rater reliability but with poor agreement between raters.
- Our study highlights the challenges (and a number of useful insights) for future trial design.

## **Background**

The reliability of rheumatologists grading digital ulcers (DUs) in patients with systemic sclerosis (SSc) has been reported to be poor to moderate at best (1–3), which is a major concern in the design of future clinical trials. Despite a number of drug therapies (4–7) to prevent and treat incident DUs, recurrent DUs are a major source of pain and disability in some patients with SSc (8). Therefore, there is a strong unmet clinical need to develop effective therapies to reduce the burden of DU disease in patients with SSc, underpinning the need for high quality clinical trials.

Recent multi-centre clinical trials of drug therapies for DUs have used different definitions for DUs in their study design. In general, previous definitions (5,7,9,10) have included a loss of surface epithelisation and with a discernible depth. Many studies have only included those DUs which occur on the fingertips, as these are considered ‘ischaemic’ and therefore presumably most likely to respond to vascular therapies, excluding those which occur over the extensor aspect of the fingers. There are also a number of digital lesions (e.g. pitting scars and fissures) which are common in patients with SSc, and can be very challenging to distinguish from DUs. Furthermore, the inclusion of DUs in the current American College of Rheumatology/ European League Against Rheumatism SSc classification criteria (11) highlights the importance of accurate recognition of DUs in patients with SSc.

Against this background, a United Kingdom Scleroderma Study Group (UKSSG) working group was convened to develop new proposed DU definitions to improve the reliability of DU grading between raters. The aim of this study was to examine the reliability of the proposed DU definitions amongst UK clinicians with an interest in SSc.

## **Patients and methods**

### UKSSG working group

Under the auspices of the UKSSG, a working group was assembled consisting of 8 UK rheumatologists with an interest in SSc, one European rheumatologist (contributing live remotely), one dermatologist, a hand surgeon and a rheumatology specialist nurse. A statistician with extensive experience of reliability research and two patients with SSc with a history of DUs were also members of the working group.

### Consensus meeting and derivation of proposed DU definitions

A DU consensus meeting was convened at the University of Manchester on the 24th November 2015. Previous DU definitions and issues around the challenges of DU grading were discussed. A key issue that emerged from the meeting was that different outcome measures might be required for preventive studies versus those studies investigating treatments for DUs. After the meeting, based upon the discussions, proposed DU definitions were drafted. These were then sent to the members of the working group for comment before being finalised. The final definitions for 'no ulcer', 'healed ulcer' and 'DU' are presented in Table 1.

### Study design and participants

Eighty clinical images of a range of digital lesions (mainly DUs) from our previous reliability study (3) were used to conduct the grading exercise. These were prospectively selected by two individuals (MH and AH) to encompass the range of digital lesions observed in patients with SSc-spectrum disorders. A gangrenous digit was specifically included, as this is a controversial issue in the definition of DUs. As previously described (3) a clinical photograph of the digital

lesion was obtained by a trained medical photographer; with a 1 cm graded scale positioned in close proximity to the lesion, to give raters an indication of the lesion size. Patient and lesion characteristics have been previously reported (3). The study was approved by the National Research Ethics Committee East of England - Hatfield, and all patients provided signed informed consent.

A new custom-built, secure web-based interface was constructed to both display and record the grading of the clinical images. All clinical members of the UKSSG (including those within the working group) were invited to participate in the web-based study. On the entry screen the proposed DU definitions were presented without any exemplar images. The definitions could be recalled for review throughout the duration of the study. Each rater graded 90 images: 80 unique and then 10 repeated images (from the first 50) to allow an assessment of intra-rater reliability. The unique images were displayed in a randomised order to each rater. Raters graded each image according to the proposed definitions on a 3-point ordinal scale of severity: either 'no ulcer' (0), 'healed ulcer' (1) or 'DU' (2). Raters had only one opportunity to score the image before choosing to move onto the next image.

### Statistical analysis

The reliability of categorical data (here 'no ulcer', 'healed ulcer' and 'DU') can be assessed by using kappa coefficients, which calculate the level of agreement between raters. Where the scale is ordered a weighted kappa coefficient (which is also an intraclass coefficient) is used. Similarly to our previous studies (1,3), overall intrarater reliability was assessed using a weighted kappa coefficient with quadratic weights, and is also presented as the mean of the individual graders. One-way ANOVA was used to assess overall inter-rater reliability. Data was dichotomised by adjoining adjacent categories which could be considered as applicable to 'preventative studies' (i.e. no ulcer vs healed ulcer and DU) and 'studies of treatments for DUs' (i.e. no ulcer and healed ulcer vs DU). It has been suggested that the kappa can be interpreted as no better than chance alone (<0), (0.01–0.20), fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80), almost perfect (0.81–0.99), and perfect (1) agreement between raters (12). All statistical analyses on the data were performed using STATA, version 13.

## **Results**

23 UK clinicians (raters): 18 rheumatologists, three dermatologists, one orthopaedic hand surgeon and one specialist rheumatology nurse successfully completed the study. A total of 2070 (1840 unique + 230 repeat) image gradings were obtained.

### **Intra-rater reliability**

The overall intra-rater reliability was high. The overall weighted kappa ( $\kappa$ ) coefficient was 0.71 (95% CI = 0.63 – 0.79) and was moderate when averaged across raters ( $\kappa = 0.55$ , SD = 0.31). Intra-rater reliability was high for both the dichotomised analyses of no ulcer vs healed ulcer and DU ( $\kappa = 0.70$ , 95% CI = 0.62 – 0.79) and no ulcer and healed ulcer vs DU ( $\kappa = 0.77$ , 95% CI = 0.67 – 0.86).

### **Inter-rater reliability**

The overall inter-rater reliability was poor ( $\kappa = 0.15$ , 95% CI = 0.10 – 0.21). Inter-rater reliability was fair for the dichotomised analyses of no ulcer vs healed ulcer or DU ( $\kappa = 0.25$ , 95% CI = 0.19 – 0.31) and moderate for no ulcer and healed ulcer vs DU ( $\kappa = 0.41$ , 95% CI = 0.33 – 0.49).

### **Definition exemplar images**

Figure 1 illustrates a number of exemplar images of the proposed DU definitions. These were selected as described in the figure legend.

## **Discussion**

The key finding of our study is that although our proposed UKSSG DU definitions had good intra-rater reliability, the agreement between raters was poor. This further confirms the urgent need to develop reliable methods to measure DUs as outcome measure in future multi-centre studies involving multiple raters.

The overall inter-rater reliability was lower than previously reported (1–3). In our previous web-based DU reliability study (3), the addition of ‘real world’ (e.g. pain and discharge) clinical contextual information did not significantly increase the inter-rater reliability ( $\kappa = 0.32$  without or  $0.36$  without the contextual information). The poor agreement between raters may be related due to the intrinsic properties/performance of our proposed definitions and/or due to differences in rater opinion. The high intra-rater reliability further confirms the importance for the *same* individual to assess patients in clinical trials, to minimise the impact of differences in opinion between raters.

A key strength of our study is that the definitions were developed by a diverse working group including colleagues from related specialities (dermatology, orthopaedic hand surgery and rheumatology specialist nursing), and with patient representation. Furthermore, a large (>2000) images were graded, which allowed analysis of rater reliability.

Our UKSSG definitions can be considered as complementary to the recently proposed definitions by the World Scleroderma Foundation (WSF) (13). Both (sets of) definitions feature a loss of depth/epithelium as a central feature of DUs. In addition, both recognise that DUs are often covered by an overlying crust or eschar, and therefore a caveat is added to both, that if debridement would likely confirm a DU, then the lesion should be classified as a DU. Neither set of definitions included an ‘unclassifiable’ category, as this was not felt to be helpful in the grading of DUs. In our definitions we chose to encompass the spectrum of DU disease, including extensor DUs and those which occur in relation to subcutaneous calcinosis.

There is a strong rationale to develop *objective* methods to measure DUs in future clinical trials, for example by using ultrasound to assess surface area and/or depth. In a pilot study in 10 patients with SSc with 15 DUs, high-frequency ultrasound was found to be a feasible method to measure a range of SSc-related DUs (14). Similarly, Sulliman et al (15) reported (currently only in abstract form), successful measurement of SSc-related DUs by musculoskeletal ultrasound. In a recent study (16), digital planimetry by free hand or fitting a semi-eclipse was found to be a reliable method to measure DU surface area, with good agreement between the



two techniques. Baron et al (2) reported moderate intra- (0.57) and inter-rater (0.48) reliability for the measurement of DU by surface area, in their study using digital callipers.

At present, assessment of treatment efficacy in clinical trials is primarily based upon clinician opinion alone: patient opinion has been less studied. In a recent reliability study, the agreement between individual patients and rheumatologists was poor without and with the clinical context (0.19 & 0.28, respectively) (3). Therefore, there is a major unmet need to develop a novel patient reported outcome (PRO) measure which captures the multifaceted patient experience of DUs to facilitate future clinical trials.

Our study has a number of important considerations. This was a web-based study and it could be argued that there is an important difference between assessing clinical photographs and physical examination of lesions. The inter-rater reliability of clinicians physically assessing digital lesions using the WSF definition was reported to be 0.5 (13) but as highlighted by the authors, both the number of patients assessed and clinicians who graded the lesions in this exercise was small (both  $n = 7$ ). Although it could be argued that in the present study the high intra-rater reliability could be due to rater recall, each rater graded a large number of images, and the 10 repeat images were drawn from the first 50 assessed.

Our study highlights a number of important lessons for the design of future SSc clinical trials, including the definition of DUs. Firstly, the development of our proposed DU definitions benefited from a diverse multi-disciplinary working group, including patient representation. Secondly, different outcome measures may be needed in preventative studies compared to those of treatments for DUs. In our study, inter-rater reliability was found to be highest in the context of 'studies of treatments for DUs' compared to 'preventative studies' and overall reliability, which could indicate raters find this classification of lesions useful in this context. Thirdly, the role of training to improve the reliability of rater grading warrants investigation. Finally, a number of images had very high or perfect rater agreement, and future studies should consider the production of an 'atlas' of exemplar images to exemplify DU definitions.

In conclusion, although our proposed DU definitions had high intra-rater reliability, the agreement between raters was poor, confirming the need to improve DUs as outcome measure in clinical trials. Our study provides a number of invaluable insights for the study design of future DU clinical trials. Future research is needed to improve the reliability of DUs as an outcome measure, including the role for objective measurement techniques and PROs.

### **Acknowledgement**

We are grateful to all the clinicians (in alphabetical order) who participated in the study: Yasmeen Ahmad, Marina Anderson, Eileen Baidem, Theresa Barnes, Maya Buch, Francesco Del Galdo, Emma Derrett-Smith, Frances Hall, Neil McHugh, Simon Meggitt, Catherine H Orteu, Claire Pain, Muditha Samaranayaka, Anita Smyth, Douglas Veale. We acknowledge the contribution made by Vasiliki Tsami in the construction of the web-based interface. We would also like to thank Mr. Steve Cottrell and colleagues from the Medical Illustration, Salford Royal NHS Foundation Trust, for collecting the images.

## **References**

1. Herrick AL, Roberts C, Tracey A, et al. Lack of agreement between rheumatologists in defining digital ulceration in systemic sclerosis. *Arthritis Rheum.* 2009;60(3):878–82.
2. Baron M, Chung L, Gyger G, et al. Consensus opinion of a North American Working Group regarding the classification of digital ulcers in systemic sclerosis. *Clin Rheumatol* 2014;33(2):207–14.
3. Hughes M, Roberts C, Tracey A, Dinsdale G, Murray A, Herrick AL. Does the clinical context improve the reliability of rheumatologists grading digital ulcers in systemic sclerosis? *Arthritis Care Res (Hoboken)* 2016;68(9):1340–5.
4. Wigley FM, Wise RA, Seibold JR, et al. Intravenous iloprost infusion in patients with Raynaud phenomenon secondary to systemic sclerosis. A multicenter, placebo-controlled, double-blind study. *Ann Intern Med* 1994;120(3):199–206.
5. Matucci-Cerinic M, Denton CP, et al. Bosentan treatment of digital ulcers related to systemic sclerosis: results from the RAPIDS-2 randomised, double-blind, placebo-controlled trial. *Ann Rheum Dis* 2011;70(1):32–8.
6. Hughes M, Ong VH, Anderson ME, et al. Consensus best practice pathway of the UK Scleroderma Study Group: digital vasculopathy in systemic sclerosis. *Rheumatology* 2015;54(11):2015–24.

7. Hachulla E, Hatron PY, Carpentier P, et al. Efficacy of sildenafil on ischaemic digital ulcer healing in systemic sclerosis: the placebo-controlled SEDUCE study. *Ann Rheum Dis* 2016;75(6):1009–15.
8. Matucci-Cerinic M, Krieg T, Guillevin L, et al. Elucidating the burden of recurrent and chronic digital ulcers in systemic sclerosis: long-term results from the DUO Registry. *Ann Rheum Dis* 2016;75(10):1770-6.
9. Gliddon AE, Doré CJ, Black CM, et al. Prevention of vascular damage in scleroderma and autoimmune Raynaud's phenomenon: a multicenter, randomized, double-blind, placebo-controlled trial of the angiotensin-converting enzyme inhibitor quinapril. *Arthritis Rheum* 2007;56(11):3837–46.
10. Khanna D, Denton CP, Merkel PA, et al. Effect of macitentan on the development of new ischemic digital ulcers in patients with systemic sclerosis: DUAL-1 and DUAL-2 randomized clinical trials. *JAMA* 2016;315(18):1975–88.
11. van den Hoogen F, Khanna D, Fransen J, et al. 2013 classification criteria for systemic sclerosis: an American college of rheumatology/European league against rheumatism collaborative initiative. *Ann Rheum Dis* 2013;72(11):1747–55.
12. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. *Fam Med* 2005;37(5):360–3.
13. Suliman YA, Bruni C, Johnson SR, et al. Defining skin ulcers in systemic sclerosis: systematic literature review and proposed World Scleroderma Foundation (WSF) definition. *J Scleroderma and Relat Disord* 2017;2(2):115-120.
14. Hughes M, Moore T, Manning J, Dinsdale G, Herrick AL, Murray A. A pilot study using high-frequency ultrasound to measure digital ulcers: a possible outcome measure in systemic sclerosis clinical trials? *Clin Exp Rheumatol* 2017 [Epub ahead of print].
15. Suliman, Raganath V, Kafaja S, Furst D. Novel use of musculoskeletal ultrasound (MSUS) to measure ulcers in the skin of systemic sclerosis (SSc) patients. *Arthritis Rheum* 2015;67(Suppl 1):2988.
16. Simpson V, Hughes M, Wilkinson J, Herrick AL, Dinsdale G. Quantifying digital ulcers in systemic sclerosis: Reliability of digital planimetry in measuring lesion size. *Arthritis Care Res (Hoboken)* 2017 [Epub ahead of print].

Digital ulcer	<p>A lesion (on the finger on or distal to the metacarpophalangeal joint) with loss of surface epithelisation and a visually discernible depth. The ulcer bed is often wet in appearance with surface slough.</p> <p>The peri-lesional skin surrounding digital ulcers is not uncommonly erythematous and/or macerated (including in the absence of superadded infection). Patients often report pain (which may be severe) associated with digital ulcers. Digital ulcers often have an overlying scab (eschar) and if there is a high index of suspicion of an underlying digital ulcer, then the lesion should be classified as such. Common sites for digital ulcers include the fingertips and over the extensor (dorsal) aspects of the hands, and in relation to subcutaneous calcinosis. Less often digital ulcers may occur at other sites on the hands (e.g. over the lateral aspects of the digits and at the base of the nail).</p>
Healed ulcer	A lesion with complete surface epithelisation (otherwise the lesion would be classified as a ‘digital ulcer’).
No ulcer	Any lesion which does not fulfil the definitions of either a ‘digital ulcer’ or ‘healed ulcer’ including (but not limited) to: digital pitting scars, hyperkeratosis, and fissures.

Table 1: Proposed UKSSG working group DU definitions.

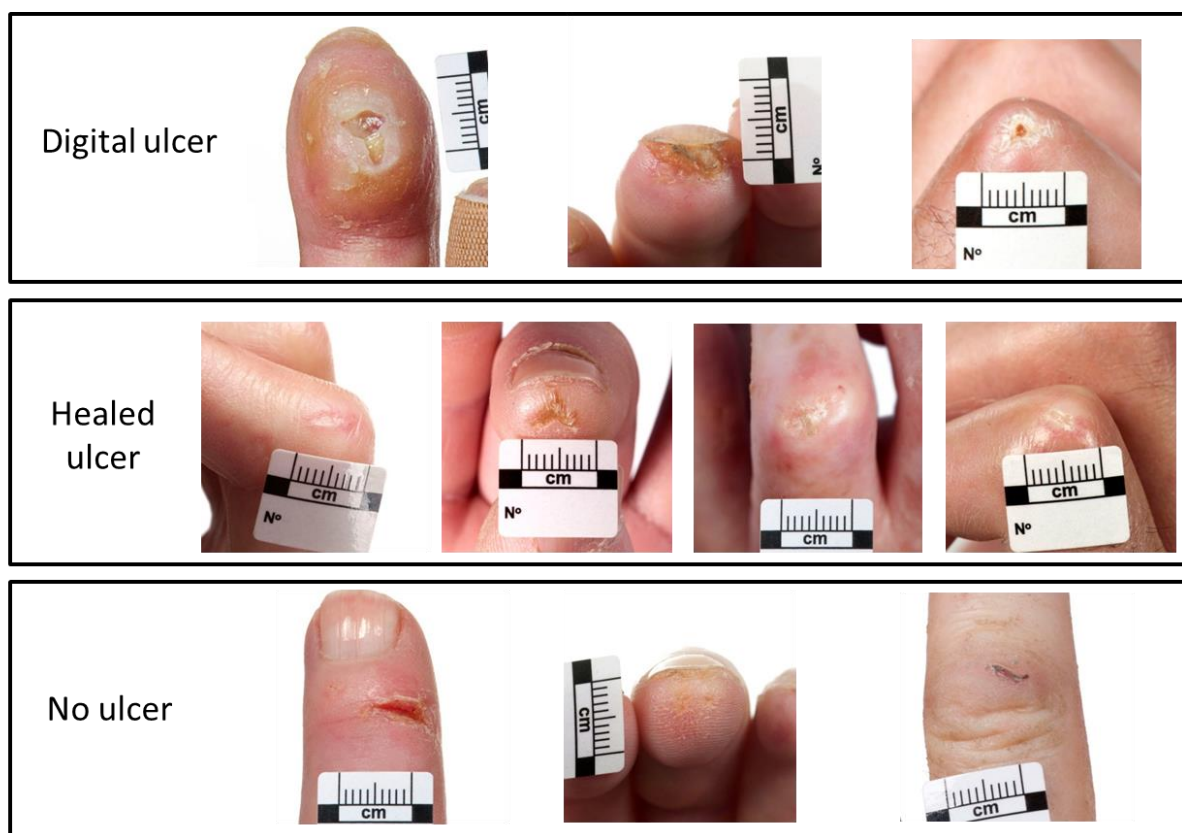


Figure 1: Exemplar images of the proposed UKSSG working group DU definitions. All the DU (top row) images had perfect (n=23/23) agreement between raters. Rater agreement (n) was lower for healed ulcer (n=18/20) and no ulcer (n=17/23).