Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/burns

Investigating the intra- and inter-rater reliability of a panel of subjective and objective burn scar measurement tools[☆]

K.C. Lee[a], A. Bamford[b], F. Gardiner[b], A. Agovino[a], B. ter Horst[a], J. Bishop[c,1], A. Sitch[d,1], L. Grover[e], A. Logan[f], N.S. Moiemen[a,*]

[a] Scar Free Foundation Centre for Burns Research, University Hospital Birmingham NHS Foundation Trust, Queen Elizabeth Hospital Birmingham, Mindelsohn Way, Edgbaston, Birmingham, B15 2WB, UK

[b] University Hospital Birmingham NHS Foundation Trust, Queen Elizabeth Hospital Birmingham, Mindelsohn Way, Edgbaston, Birmingham, B15 2WB, UK

[c] Birmingham Clinical Trials Unit (BCTU), Institute of Applied Health Research, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

[d] Test Evaluation Research Group, Institute of Applied Health Research, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

[e] School of Chemical Engineering, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

[f] Institute of Inflammation and Ageing, College of Medical and Dental Sciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

ARTICLE INFO

Article history:
Accepted 7 February 2019

Keywords:
Burn scar
Objective measurements
Reliability

ABSTRACT

Background: Research into the treatment of hypertrophic burn scar is hampered by the variability and subjectivity of existing outcome measures. This study aims to measure the inter- and intra-rater reliability of a panel of subjective and objective burn scar measurement tools.

Methods: Three independent assessors evaluated 55 scar and normal skin sites using subjective (modified Vancouver Scar Scale [mVSS] & Patient and Observer Scar Assessment Scale [POSAS]) and objective tools. The intra-class correlation coefficient was utilised to measure reliability (acceptable when >0.70). Patient satisfaction with the different tools and scar parameter importance were assessed via questionnaires.

Results: The inter-rater reliabilities of the mVSS and POSAS were below the acceptable limit. For erythema and pigmentation, all of the Scanoskin and DSM II measures (except the b* value) had acceptable to excellent intra and inter-rater reliability. The Dermascan ultrasound (dermal thickness, intensity) had excellent intra- and inter-rater reliability (>0.90). The Cutometer R0 (firmness) had acceptable reliability but not R2 (gross elasticity). All objective measurement tools had good overall satisfaction scores. Patients rated scar related pain and itch as more important compared to appearance although this finding was not sustained when corrected for multiple comparisons.

[☆] Name of the department or institution to which the work should be attributed: Scar Free Foundation Centre for Burns Research, University Hospital Birmingham NHS Foundation Trust, Queen Elizabeth Hospital Birmingham.

[*] *Corresponding author.*
E-mail address: nmoiemen@aol.com (N.S. Moiemen).

[1] AS and JB were supported by the NIHR Birmingham Biomedical Research Centre at the University Hospitals Birmingham NHS Foundation Trust and the University of Birmingham.

https://doi.org/10.1016/j.burns.2019.02.002

Conclusion: The objective scar measures demonstrated acceptable to excellent intra- and inter-rater reliability and performed better than the subjective scar scales.

1. Introduction

Research into burn scar treatment has increased in recent years as the importance of post-recovery quality of life for burn victims is being recognised. However, research into scar treatments is hampered by the highly variable nature of burn scars as they are influenced by both injury factors (e.g. cause, depth of burn, infection, type and timing of surgical treatment and time to healing [1]), surgical factors (e.g. number of procedures, type of skin grafts [2]) and patient factors (e.g. gender, ethnicity, age, anatomical site, comorbidities [2,3]).

Additional variability is introduced by the subjective methods of evaluating scars which is prevalent in clinical practice and research. Subjective scores such as the Vancouver Scar Scale (VSS) and Patient and Observer Scar Assessment Scale (POSAS) are commonly used [4] as they are low-cost, quick and easy methods as well as validated and widely published. The reproducibility of these subjective tools however are highly reliant on the users; if utilised by trained clinicians who have similar, agreed opinions on scars, the reliability can be very good [5]. However this is rarely the case outside of research where different clinicians' judgement of scars can be vastly different depending on their previous experience with scars and patients' perception of their own scars can be similarly influenced by a myriad of factors such as the trauma of the injuring event and the treatment process, previous experiences, psychological state and visibility of the scar [6]. Disagreement between scar ratings between clinicians and patients is common [7] and differences can indicate the presence of psychological distress [6].

The use of objective measurement methods can aid in reducing this variability. Objective measurements are methods of quantifying a property of the scar that is minimally influenced by the user, patient and the innate random and systematic errors of the instrument itself, i.e. its reliability [8]. Scar assessment consists of the measurement of multiple components including colour (erythema and pigmentation), pliability, thickness, and irregularity. A small number of devices have been adapted from the cosmetics industry for burn scar measurements as there is an increasing awareness of the importance of reproducible and objective scar measurements for both clinical practice and research. However there is currently no consensus on the most suitable tools for measurement of the different aspects of scars due to the scarcity of scientific studies in these instruments. This study is one of the first steps in establishing a panel of objective scar measurement devices. In order to inform the choice of the devices as well as to understand the limitations of the different measurement methods. Furthermore, the study aims to measure the reliability of both subjective measurement tools as well as a panel of objective measurement devices. It is hoped that the results in this study will inform future scar research and improve the accuracy, objectiveness and reproducibility of measured scar outcomes in these studies.

2. Methods

The study was conducted in the Wellcome Trust Clinical Research Facility at the Queen Elizabeth Hospital (QEHB), Birmingham, United Kingdom. This study was approved by the South Birmingham National Research Ethics Service (NRES) Committee West Midlands (REC reference 15/WM/0378) as well as the Research and Development Governance of the University Hospitals of Birmingham.

2.1. Study population

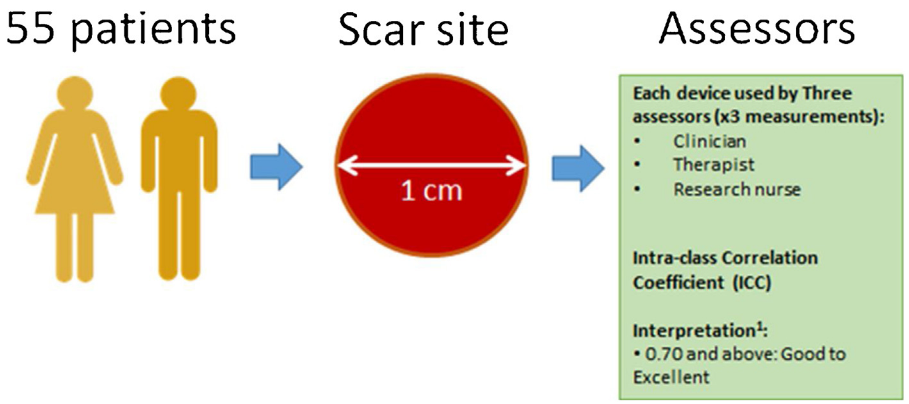
Fifty-five adult patients who have been treated at QEHB were invited to participate in this study, and participation was voluntary. Subjects were included in the study if they met the following inclusion criteria: age 18 years and above, hypertrophic but non-keloid burn scar, scar aged more than 3 months (calculated from time of 95% healing,) or have had a skin-grafted area, or a burn that has had delayed healing (>2 weeks), scar size of at least 10 cm^2. We excluded patients if they had other pathological skin conditions, chronic steroid use and scar areas on the genitalia or face. This was done to minimise the interference of pathological skin on device readings for example excessive skin flaking in psoriasis may clog the suction based Cutometer, and abnormally rigid skin (e.g. scleroderma) or loose skin (e.g. Ehler Danlos) will give abnormal control skin values which will make analysis of scar to normal skin ratios and correlations difficult due to outlier values.

Potential subjects were identified from outpatient scar management therapy lists by clinicians and therapists or from multi-disciplinary team meetings and ward rounds. They were then contacted by a member of the research team; either by telephone or during routine clinical appointments and a brief explanation of the trial was given. If they agreed to participate in the study, a Patient Information Leaflet (PIL) was then sent to the participant.

2.2. Measurement procedures

The study was a prospective, non-blinded single-arm observational study using three independent assessors. A graphical overview of the study design and pathway for the scar assessment study day is shown in Fig. 1.

All study participants had a single scar site which was deemed the worse by both patient and clinician chosen. The clinicians were largely guided by the patient in the selection of the scar site. The clinicians' role was to ensure that the selected scar site met the protocol of the study (e.g. not on the face or genitalia areas) and could be feasibly measured by the



Each device used by Three assessors (x3 measurements):

- Clinician
- Therapist
- Research nurse

Intra-class Correlation Coefficient (ICC)

Interpretation[1]:
- 0.70 and above: Good to Excellent

Fig. 1 – Graphical overview of study design.

devices. If there were disagreements, a consensus was sought between the patient and three clinicians. Within this scar site, a 3 × 3 cm area was then selected and marked. One 1 cm circle site was then selected and marked (with a stencil and marker pen) for evaluation within this area on each participant. One site of normal skin (1 cm circle area) that is corresponding to the same anatomical site (contralateral site, or adjacent anatomic site) was also chosen for measurements. Similarly, the site was selected and marked for evaluation within this area on each subject. The 1 cm size of the marking was chosen as a balance between being bigger than the aperture sizes of the devices (0.6–0.8 cm) and small enough to minimise variability due to relocation.

If the patients were wearing pressure garments or gels/moisturisers, they were asked to remove them at least 20 min before their appointment. Measurements were conducted in the same temperature controlled room (22+/–1 °C) with the patient lying in the same position for each consecutive measurement. The temperature and humidity of the room was measured and monitored.

This scar site (as well as a normal skin site which was contralateral or adjacent to the measured scar site) was measured by 3 different assessors 3 times over a period of one day with each of the devices in the panel of objective scar measurement tools, a total of 9 measurements per device per site. As the scars were measured in the same day, subjective scar scales (mVSS and POSAS) were only completed once by each of the three assessors and patient. For the objective measurement devices, the normal skin sites (contralateral or adjacent) of the patients served as the control groups.

2.3. *Raters*

The same three raters (a clinician, research nurse and burns occupational therapist) were used in the entirety of the study.

The burns research fellow, nurse and therapist have formal training and experience in the methodology of scar assessments. The burns research fellow prior to conducting the clinical trial, which is part of his PhD thesis, had in-depth training in all subjective and objective scar assessments. The research nurse is a senior burn nurse who was involved in scar assessment objective and subjective for years prior to the trial. The occupational therapist had 10 years' experience in scar assessment and management of post burn hypertrophic scarring including the objective methods used in the trial.

In addition to this, all three raters were provided with comprehensive training in the use of the objective scar measurement devices under the supervision of the company representatives of the respective devices and prior to the study commencing, training sessions which involved the raters using the devices on actual patients with burn scars were run to allow the raters to familiarise themselves further with the devices and also the running of the study.

2.4. *Measurement tools*

In a previously published study, a review of the current literature on methods of objective scar measurements was performed [9], and three parameters of scar assessment that are the most commonly measured as well as the tools required to measure them have been identified: pliability, scar thickness and colour. These can be divided into subjective and objective measurement tools.

2.4.1. *Subjective measurement tools*

All subjective measurement scales were completed with pen and paper when face to face with the patient.

2.4.1.1. *Modified Vancouver Scar Scale (mVSS)*. A modified version of the VSS that was adapted from the modified version used by Nedelec et al. [10,11] is used in this study (Table 8). This scale uses a numerical assessment of four skin characteristics including: Height (range, 0–4), Pliability (range, 0–4), Vascularity (range, 0–3), and Pigmentation (range, 0–3). The larger the number the worse the scar. The assessors choose a numerical value for each of these characteristics based on a comparison with normal skin.

2.4.1.2. *Patient and Observer Scar Assessment Scale (POSAS)*. The Patient and Observer Scar Assessment Scale (POSAS, version 2.0) is a subjective scar scale that consists of two parts: a Patient Scale and an Observer Scale [5]. Both scales contain six items that are scored numerically on a ten-step scale (i.e. 1–

10). A score of “1” being “no, not at all” and a score of “10” being “yes, very much”. Together they make up the ‘Total Score’ of the Patient and Observer Scale.

2.4.2. *Objective measurement tools*

2.4.2.1. *DSM II Colormeter*. The DSM II Colormeter (Cortex Technology ApS, Denmark) is a small handheld device which combines two methods of quantifying colour: narrow-band spectrophotometry (melanin, erythema) and tristimulus reflectance colorimetry in a single measurement [12]. The Colormeter consists of a probe which is made of a transparent dome which houses 2 white LED lights and a colour sensor has a skin measuring area of 4mm in diameter. Measurements were done by placing the probe over the selected areas on the scar. The probe is to be held perpendicular to the scar using minimal pressure to avoid blanching of the scar.

2.4.2.2. *Scanoskin camera*. The Scanoskin camera system (Leniomed Ltd, United Kingdom) is a new device which is a type of spectrophotometer which is a device that can measure a light beam’s intensity as a function of its colour (wavelength) [13]. The system consists of a standard DSLR camera, a polarising light filter, ring flash and Scanoskin software. The polarising light filter and ring flash produces a standard controlled lighting which illuminates the skin. Some of this light is reflected and scattered from the surface of the skin. This reflected light is captured with the DSLR camera and the raw image is then processed by the Scanoskin software which then splits it into 2 images, one for melanin (pigment found in skin) and the other for haemoglobin (pigment found in red blood cells) (Fig. 2). The image is first inverted (thus darker pixel intensity will have higher values). The Region Of Interest (ROI) is then traced using the Image-J software and then the Histogram function is used to give the mean pixel intensity.

2.4.2.3. *Dermascan 20MHz high frequency ultrasound*. The Dermascan C USB (Cortex Technology ApS, Denmark) is a high-frequency (20MHz) ultrasound scanner that enables the imaging of soft tissue at high resolution with a computer, and comes with software that allows automated skin thickness measurement [14].

In the study, a medium focus transducer was used with a 12mm wide viewing field and penetration depth of 15mm. Before measurement, a thin layer of conducting ultrasound gel is applied to the transducer and the transducer is to be held perpendicular to the scar sites to record a single echographic image for each site. Mode 4 and a gain profile of 13 were set for all scans.

All measurements were performed with an ultrasound frequency set at 1580m/s. Thickness measurements are then generated by a single researcher with the provided dedicated software (Advance Control 6 Analysis SW package, Cortex). The B-mode is utilised to analyse the Images [15]. This mode provides a two-dimensional ultrasound image display consisting of pixels with varying intensities to represent the amplitude of the returned echo signal. The thickness measured is defined as the distance between the top layer of the dermis underneath the echogenic stratum corneum and the inner layer of the dermis (in millimetres). Normal dermis is highly echogenic due to the high amount of connective tissue and collagen (Fig. 3). Scar tissue appears hypoechoic compared to normal skin, and this may be due to the increased water content of scar tissue due to aberrant proteoglycan metabolism.

2.4.2.4. *Cutometer elasticity probe*. The Cutometer (MPA 580, Courage and Khazaka GmbH, Germany) is an electronic instrument that assesses skin elasticity [16]. The probe of the device is placed over the area of measurement, which then generates a negative pressure that draws the skin into a hollow aperture in the centre of the probe and then uses a laser to estimate the amount of skin displacement.

The probe with a 6-mm diameter hollow aperture was chosen for this study as previous studies have determined it to be the most efficient size to measure the visco-elasticity properties of the dermis. For this study, mode 1 was chosen. This delivers three cycles of negative air pressure (500mbar) for 2s, followed by 2s of no pressure. Results are expressed as the means of the three measurement cycles. The most commonly reported R-parameters, R0 and R2 were used in this study. R0 describes the maximum deformation (extension) of the skin. R2 is the ratio of the final retraction and the maximum deformation.

2.5. *Statistical analyses*

All statistical analyses of the data were performed using IBM SPSS Statistics, version 23 (IBM Corporation, New York, USA) and Stata version 15. P values of <0.05 are considered to signify statistical significance.

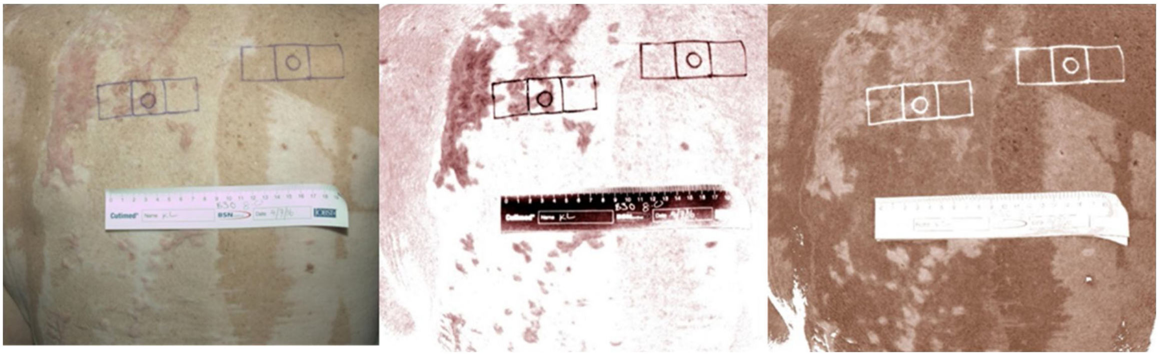


Fig. 2 – Scanoskin images: (a) original image, (b) haemoglobin image, (c) pigmentation image. Dermascan 20MHz high frequency ultrasound

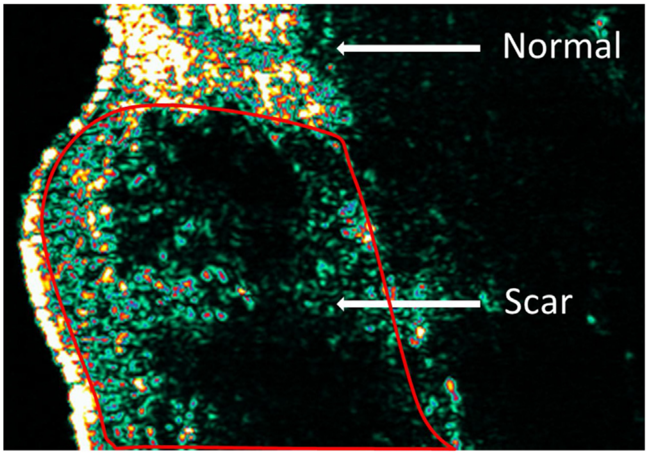


Fig. 3 – High frequency ultrasound image of scar and normal skin.

2.5.1. *Sample size*

The primary outcome was the intra and inter-reliability of a panel of objective scar measurement devices. To estimate the ICC with a 95% CI with an interval no greater than 0.2, according to the methods in Shoukri et al. [17], the required sample size was a minimum of 54 subjects. The 95% CI width of 0.2 was chosen due to a mixture of desired precision and pragmatism and allowed for an attainable sample size for the study.

Testing was conducted on 55 patients for all measures/ scales except where indicated.

2.5.2. *Reliability*

Reliability of a measurement refers to the consistency of the data when the same trait is measured by the same assessor (intra-rater reliability) or by different assessors (inter-rater reliability) with the same measurement device.

The intra- and inter-rater reliability of the objective (DSM II Colormeter, Cutometer, Dermascan, Scanoskin) and subjective measurement tools (total mVSS and POSAS Scores; sum of pliability, height, and vascularity subscales) were examined by calculating the the Intraclass Correlation Coefficients (ICC) with 95% Confidence Intervals (CI). The ICC's (2,3) were computed as per Shrout and Fleiss [18] based on the two-way random effect analysis of variance model with the absolute agreement type being selected for ICC calculations. ICC agreement and consistency were calculated by fitting a random effects model accounting for the clustering of readers within patients and reads within readers with models fitted using restricted maximum likelihood (REML) which allowed estimates of the variability at the levels of between patients, between readers and within readers to be calculated from these ICCs [19].

The two-way random effect method was selected as the raters in this study are consistent (i.e. the same three raters throughout the study) and are a sample of raters (rather than a fixed population). In addition to the ICC, the Standard Error of Mean (SEM) and the Coefficient of variation ([CV = Standard error/mean] × 100) are also reported.

The three measurements performed by each assessor for each device were used to calculate the intra-rater ICC for each assessor and then the average of the three intra-rater ICC values was calculated. For the inter-rater ICC, the mean of all three assessments for each device was calculated for each assessor and then subsequently used to calculate the ICC.

The ICC gives the proportion of variability at the between participant level (values close to 1 show less variability in the process of obtaining the measure). The ICC can be reported as a "single measure ICC" or an "average measure ICC". The single measure ICC is based on a randomly taken single measurement and is equivalent to the reliability of a measurement carried out by a single observer. The average measure ICC is based on the average measurements of three observers and thus is equivalent to the reliability of a measurement carried out by three observers. An ICC value of 0.70 was selected as the minimal threshold requirement for measurements to be deemed reliable.

As the POSAS subscales are rated from 1 to 10 (and ordered,), they can be treated as a continuous variable and hence the ICC is used to calculate the reliability. For the mVSS subscales, as the scoring is categorical and mostly ordinal in some subscales, the Fleiss kappa [20] (instead of the Cohen's kappa which is only suitable for 2 raters) and Krippendorf's alpha were reported. The reliability of the total score of the mVSS however was calculated using the ICC as it is the sum of the scores and can be viewed as a continuous variable. Similar to the ICC, a negative or low kappa/Krippendorf alpha score (≤ 0 or 0 to 0.20) indicates "no agreement" and a minimum value of 0.70 was chosen as the threshold to be deemed acceptable although it is noted in the literature that lower values (0.40–0.60) have been accepted as adequate [21,22].

2.5.3. *Patient rated scar parameter importance*

Patients were also be asked to rank the different scar parameters (surface area, thickness, colour [erythema and pigmentation], pliability, and pain/itch) in terms of importance to them, with 1 being the most important, to 6 being the least important to them. The Friedman and Wilcoxon signed-rank test was performed to examine the differences in ranking of the different scar parameters and the differences in parameters are reported with p-values.

3. Results

3.1. Study population

3.1.1. Demographics

In total, 55 participants were successfully recruited for this study between January to October 2016. The study was conducted by organising 27 outpatients clinics with 1–3 patients in each with the average time taken for each patients being about 3h. There were 37 males and 18 females, and the mean age was 46 years (range: 18–77 years). The majority of participants were Caucasian (82%). The demographics of the participants and the clinical characteristics of the scar sites chosen for evaluation are reported in Table 1.

Most of the participants were male (n=37, 67.2%) and the mean total body surface area (TBSA) burned was 16.5%. The scars were measured an average of 14.6 months after time of 95% healing (which was calculated as 7days after the last grafting procedure unless there is evidence in the patient records of delayed healing). The most commonly measured scar area was on the upper limb (41.8%) and the most common cause of scars in our cohort were flame burns (58.2%).

Table 1 – Clinical characteristics of patients recruited into study.

Clinical characteristics	n=55
Gender	
Male:female	37:18
Age	
Mean ± SD	46 ± 17.8 years
Range	18–77 years
TBSA of burn injury	
Mean ± SD	16.5 ± 18.2%
Range	0.50–60.00%
Age of scar (months after burn)	
Mean ± SD	14.6
Range	2–43 months
Aetiology	
Flame	32
Scald	16
Contact	5
Electrical	2
Location	
Upper limb	23
Lower limb	12
Chest	2
Abdomen	13
Back	5
Previous treatment of scar site	
Conservative (non-grafted)	9
Grafted	46
Fitzpatrick skin type	
I	1
II	11
III	14
IV	18
V	11
VI	0

3.2. Subjective Scar Scales

The ICC and kappa values for the mVSS and POSAS scores are presented in Table 2. The total score of the mVSS (Pliability + Height + Vascularity + Pigmentation) and the POSAS (Vascularity + Pigmentation + Thickness + Relief + Pliability + Surface area) are the sums of the individual subscales. As both the mVSS and POSAS requires the comparison of the evaluated site (scar site) with the normal skin, the ICC of these two scores for normal skin cannot be evaluated as there should be total agreement. Additionally, the intra-rater kappa and ICC values also could not be performed as each assessor only completed the subjective scores once. This was because all assessments for the patients had to be completed in the same day and the subjective scores would be heavily influenced by recall bias. Analyses of both the total scores as well as the individual subscales were performed.

For the mVSS, the kappa values for all the subscales of pliability, height, vascularity and pigmentation all fell well below the study set threshold of 0.70, with the lowest being for pigmentation (−0.02) which indicates no agreement. The single and average ICC scores for the mVSS Total score performed better than the subscales, with a value of 0.415 and 0.68 but still remained below the acceptable threshold. Even with the Alpha-Krippendorff analysis which treats the mVSS as an ordinal score (and therefore better reliability for scores with adjacent values,), all of the values are still below 0.70.

For the POSAS score, the ICC (Single) values for the Total score and the individual subscales all fall below the 0.70 threshold with the lowest score being for Surface area (0.003). The POSAS however performs better with higher reliability scores compared with the mVSS in the corresponding subscales: mVSS Pliability versus POSAS Pliability (0.149; 0.384), mVSS Height versus POSAS Thickness (0.031; 0.492); mVSS Vascularity versus POSAS Vascularity (0.241; 0.646) and mVSS Pigmentation versus POSAS Pigmentation (−0.02; 0.304). The ICC scores of the Total scores for both mVSS and POSAS were similar (0.415 versus 0.438). The removal of the subscale with the lowest reliability (Surface area) improves the ICC (Single) value of the POSAS Total score but not enough to reach the acceptable threshold (ICC Single=0.528). The ICC (Average, three assessors) for the POSAS individual subscales and Total score showed improved reliability compared to the ICC (Single) with the Vascularity, Thickness subscales and Total scores showing ICC values above the set threshold for acceptability or close to the threshold (Relief, Pliability, Overall subscales). However, other subscales (Pigmentation and Surface area) were still below the acceptable limit. Further analyses were then performed to investigate if the ICC (Average) of two assessors were equivalent or similar to three assessors, however as Table 3 shows, the ICC (Average) of three assessors for the subscales and Total score were better compared to two assessors.

Table 2 – Inter-rater reliability values for the mVSS and POSAS subscales and total scores.

	Subscale	Mean	IQR	ICC (single)	(95% CI)	Fleiss kappa[a] or ICC (average)	(95% CI)	Alpha-Krippendorff	(95% CI)	CV% (SEM)
mVSS	Pliability[a]	2.36	1.67–3.33	–	–	0.149[a]	(0.063–0.234)	0.462	0.345–0.556	35.3 (0.41)
	Height[a]	1.43	1.00-1.67	–	–	0.031[a]	(−0.74 to 0.135)	0.182	0.030–0.333	43.3 (0.36)
	Vascularity[a]	1.30	1.00-1.67	–	–	0.241[a]	(0.135–0.347)	0.450	0.319–0.567	40.8 (0.25)
	Pigmentation[a]	1.85	1.33–2.33	–	–	−0.020[a]	(−0.119 to 0.079)	0.004	−0.172 to 0.180	50.4 (0.47)
	Total	6.94	6.00-8.33	0.415	(0.249–0.578)	0.680	(0.499–0.804)	0.367	0.228–0.489	21.7 (0.82)
POSAS	Vascularity	4.60	3.00-6.00	0.646	(0.511–0.761)	0.846	(0.758–0.905)	0.645	0.554–0.719	40.9 (0.60)
	Pigmentation	4.13	3.00–5.00	0.304	(0.139–0.478)	0.568	(0.327–0.733)	0.255	0.097–0.384	32.4 (0.78)
	Thickness	4.52	3.00–4.00	0.492	(0.298–0.656)	0.744	(0.561–0.851)	0.408	0.284–0.520	33.8 (0.84)
	Relief	4.28	3.33–5.33	0.359	(0.144–0.553)	0.627	(0.335–0.788)	0.320	0.179–0.447	37.2 (0.89)
	Pliability	4.46	3.33–5.33	0.384	(0.207–0.556)	0.652	(0.439–0.790)	0.320	0.197–0.449	35.8 (0.86)
	Surface area	2.16	1.67–2.67	0.003	(−0.033 to 0.059)	0.008	(−0.106 to 0.158)	−0.385	−0.563 to 0.186	78.2 (1.04)
	Overall	4.83	3.67–6.00	0.394	(0.148–0.600)	0.661	(0.342–0.818)	0.331	0.206–0.438	33.6 (0.89)
	Total	24.14	19.00–29.50	0.438	(0.174–0.642)	0.700	(0.387–0.843)	0.478	0.366–0.582	22.1 (2.92)

[a] The pliability, height, vascularity and pigmentation reliability for the mVSS was calculated using the Fleiss kappa and alpha Krippendorff instead of ICC (number of bootstrap samples set to 1000).

Table 3 – The ICC (Average) of two assessors versus three assessors.

POSAS subscales	ICC (average)	
	Two assessors	Three assessors
Vascularity	0.78	0.85
Pigmentation	0.41	0.57
Thickness	0.65	0.74
Relief	0.54	0.63
Pliability	0.55	0.65
Surface area	−0.03	0.01
Overall	0.57	0.66
Total	0.61	0.70

3.3. *Objective devices*

3.3.1. *DSM II Colormeter: Erythema and pigmentation*

3.3.1.1. *Erythema measurements*. Erythema can be measured with the DSM II a* and the narrow band Erythema parameters. The inter- and intra-rater ICC values for the a* and narrow band Erythema parameters can be seen in Table 4.

The a* value was found to have acceptable inter- (ICC [Single, Average]=0.718, 0.884) and intra-rater (Average ICC [Single, Average]=0.731, 0.87) ICC values for scar tissue. The ICC single value for the a* measurement of normal skin however was considerably lower compared to scar tissue (ICC=0.595 versus 0.718) although the average ICC values were comparable (ICC=0.815 versus 0.884). When this is analysed in more detailed by looking at the intra-rater ICC values for the individual assessors, it can be seen that this reduction in ICC is likely to be due to the reduced reliability of the measurements by Assessor 2 (a* ICC value [Single]=0.508 versus 0.915 for Assessor 1 and 0.952 for Assessor 3). When the ICC measurements were analysed in pairs rather than all three assessors, the lowest ICC value was found between Assessors 2 and 3 (ICC [Single]=0.399).

The a* ICC agreement for scar tissue was calculated to be 0.571 (95% CI 0.457, 0.685) and for normal skin the ICC was lower at 0.474 (0.507, 0.739).

The narrow band Erythema value was also found to have acceptable inter- (ICC [Single, Average]=0.777, 0.912) and intra-rater ICC values (Average ICC [Single, Average]=0.74, 0.87) for scar tissue. Similar to the a* value for normal skin, the ICC of the Erythema measurement for normal skin (ICC [Single] =0.647) was lower than that of scar tissue (ICC [Single]=0.777). As with the a* value, when analysed in pairs, the lowest ICC value was found between Assessor 2 and 3 (ICC [Single]=0.450). For the erythema measure the ICC agreement for scar tissue was 0.622 (95% CI=0.517–0.728) and for normal skin was again lower at 0.510 (95% CI=0.386–0.634).

For both measures and skin types, the CV was greatest at the between patient level; however, the within reader CV was larger than the between reader CV.

Of the two erythema measures, both the a* and narrow band Erythema measure have similar inter- and intra-rater reliability values as well as %CV, thus both are recommended. However for erythema measurements on normal skin, more than 1 measurement should be taken to improve reliability.

Table 4 – Inter-rater ICC values for the erythema measures of the DSM II Colormeter for scar tissue and normal skin.

Measure	Scar tissue		Normal skin	
	a*	Erythema	a*	Erythema
Mean	16.12	15.02	13.28	10.30
IQR	14.03–18.02	12.84–17.44	10.79–15.10	7.26–13.20
%CV (patient)	19.0	21.6	24.30	32.0
Inter-rater ICC				
Single ICC (95% CI)	0.718 (0.601–0.813)	0.777 (0.678–0.854)	0.595 (0.424–0.732)	0.647 (0.488–0.769)
Average ICC (95% CI)	0.884 (0.819–0.929)	0.912 (0.863–0.946)	0.815 (0.689–0.891)	0.846 (0.741–0.909)
%CV (SEM)	9.65 (0.83)	9.45(0.73)	13.77 (1.09)	14.5 (0.90)
Intra-rater ICC				
Single ICC (95% CI)	0.731 (0.233–0.949)	0.737 (0.260–0.938)	0.792 (0.352–0.970)	0.792 (0.283–0.985)
Average ICC (95% CI)	0.87 (0.477–0.983)	0.87 (0.477–0.978)	0.90 (0.620–0.990)	0.89 (0.542–0.995)
%CV (SEM)	9.83 (0.87)	10.03 (0.84)	9.43 (0.76)	8.93 (0.58)
Agreement and consistency				
ICC agreement (95% CI)	0.571 (0.457–0.685)	0.622 (0.517–0.728)	0.474 (0.343–0.604)	0.510 (0.386–0.634)
ICC consistency (95% CI)	0.649 (0.546–0.753)	0.677 (0.580–0.773)	0.623 (0.507–0.739)	0.627 (0.516–0.738)

3.3.1.2. Pigmentation measurements. Pigmentation can be measured with the DSM II L*, b* and the narrow band Melanin parameters. The inter- and intra-rater ICC values for the L*, b* and narrow band Melanin parameters can be seen in Table 5.

The L* value was found to have good inter- (ICC [Single, Average]=0.942, 0.980) and intra-rater (Average ICC [Single, Average]=0.90, 0.97) ICC values for scar tissue. The L* parameter had an ICC agreement of 0.882 (95% CI: 0.839–0.925). The inter-and intra-rater for the L* value for normal skin, although lower, were also all above the set threshold of 0.70 (ICC Single or Average=0.806–0.960).

The narrow band Melanin value was also found to have good inter- (ICC [Single, Average]=0.930, 0.975) and intra-rater (Average ICC [Single, Average]=0.93, 0.97) ICC values for scar tissue. For the melanin measure the ICC agreement was similar with 88.4% of the variability at the between-participant level (ICC=0.884 (0.841, 0.928)).The inter-and intra-rater for the Melanin value for normal skin were also similarly high with inter-rater values of 0.836 (ICC Single) and 0.939 (ICC Average); and intra-rater values of 0.90 (ICC Single) and 0.96 (ICC Average).

The b* value however performed the worst out of the three objective pigmentation measures. For the measures of L* and melanin the largest CV was between patients, however for the measure of b* the largest CV was at the within reader level. The b* value for scar tissue had a low inter-rater ICC (Single) value of 0.525, but acceptable ICC (Average) value of 0.768 for scar tissue, although this is with a high %CV of 32.58%. The b* intra-rater ICC (Single) value for scar tissue was just below the threshold (ICC Single=0.62) but the ICC (Average) was good (ICC Average=0.80). Both the ICC (Single) and ICC (Average)

Table 5 – Inter-rater and intra-rater ICC values of the pigmentation measures for the DSM II Colormeter for scar tissue and normal skin.

Measure	Scar tissue			Normal skin		
	L*	b*	Melanin	L*	b*	Melanin
Mean	27.07	6.61	46.36	37.67	9.97	35.80
IQR	22.36-32.41	3.95–8.69	39.63–49.97	37.75–43.34	7.59–12.66	30.96–38.15
%CV (patient)	26.0	42.1	20.6	18.1	24.1	20.0
Inter-rater ICC						
Single ICC (95% CI)	0.942 (0.908–0.964)	0.525 (0.370–0.666)	0.930 (0.893–0.956)	0.806 (0.576–0.903)	0.351 (0.178–0.523)	0.836 (0.677–0.909)
Average ICC (95% CI)	0.980 (0.967–0.988)	0.768 (0.638–0.857)	0.975 (0.962–0.985)	0.926 (0.803–0.965)	0.618 (0.393–0.767)	0.939 (0.873–0.968)
%CV (SEM)	6.14 (0.88)	32.58 (1.09)	4.06 (1.13)	8.10 (1.63)	32.19(1.60)	6.69 (1.43)
Intra-rater ICC						
Single ICC (95% CI)	0.90 (0.774–0.963)	0.64 (0.202–0.879)	0.93 (0.851–0.992)	0.90 (0.729–0.985)	0.62 (0.152–0.930)	0.90 (0.771–0.984)
Average ICC (95% CI)	0.97 (0.911–0.987)	0.82 (0.431–0.956)	0.97 (0.945–0.992)	0.96 (0.890–0.995)	0.80 (0.350–0.976)	0.96 (0.910–0.995)
%CV (SEM)	6.87 (1.01)	59.84 (1.09)	4.42 (1.05)	5.20 (1.03)	−63.88 (1.18)	5.00 (3.12)
Agreement and consistency						
ICC agreement (95% CI)	0.882 (0.839–0.925)	0.362 (0.237–0.487)	0.884 (0.841–0.928)	0.749 (0.662–0.837)	0.220 (0.096–0.345)	0.780 (0.702–0.858)
ICC consistency (95% CI)	0.903 (0.867–0.939)	0.437 (0.309–0.565)	0.924 (0.894–0.953)	0.886 (0.841–0.930)	0.313 (0.167–0.460)	0.889 (0.847–0.932)

values of the b* measure were below the acceptable threshold for normal skin. The b* intra-rater ICC (Single) value for normal skin was just below the threshold (ICC Single=0.62) but the ICC (Average) was good (ICC Average=0.80).

3.3.2. *Scanoskin camera: Erythema and pigmentation*

Due to technical difficulties and errors in the flash power (resulting in photo underexposure), only 31 of the patients had Scanoskin camera photos which were analysable (B25 to B55). The inter- and intra-rater ICC values for the Scanoskin parameters can be seen in Table 6.

The Scanoskin erythema measure had very high reliability values for both inter- (ICC [Single, Average]=0.969, 0.989) and intra-rater (ICC[Single, Average]=0.995, 0.998) ICC values for scar tissue. The Scanoskin erythema measure for normal skin also had slightly lower but similarly high inter- (ICC [Single, Average]=0.926, 0.974) and intra-rater (ICC[Single, Average] =0.985, 0.995) ICC values.

The same trend was seen with the Scanoskin pigmentation measure, with analyses showing high inter- (ICC [Single, Average]=0.972, 0.991) and intra-rater (ICC [Single, Average] =0.989, 0.996) ICC values for scar tissue as well as normal skin, inter-rater (ICC [Single, Average]=0.957, 0.985); intra-rater (ICC [Single, Average]=0.994, 0.997).

The ICC agreement estimates for erythema and pigmentation were very high for the measurements of scar tissue, 0.966 (95% CI: 0.945–0.986) and 0.965 (95% CI: 0.946–0.985) respectively. The ICC estimates when using the normal skin were similar. The CV at the between patient level was largest for erythema and pigmentation measures for both skin types and the within reader CV was the smallest.

3.3.3. *Dermascan ultrasound: dermal thickness and intensity*

The inter- and intra-rater ICC values for the Dermascan ultrasound dermal thickness and intensity measures are shown in Table 7.

Dermascan measured dermal thickness had good inter- (ICC [Single, Average]=0.957, 0.985) and intra-rater (ICC [Single, Average]=0.951, 0.983) for scar tissue. The dermal thickness measure for normal skin also had similarly high inter- (ICC [Single, Average]=0.967, 0.989) and intra-rater (ICC [Single, Average]=0.948, 0.982).The Dermascan measured dermal intensity likewise had good inter- (ICC [Single, Average] =0.918, 0.971) and intra-rater (ICC [Single, Average]=0.928, 0.937) for scar tissue. The dermal intensity measure for normal skin also had similarly high inter- (ICC [Single, Average]=0.863, 0.950) and intra-rater (ICC [Single, Average]=0.931, 0.976).

Although both Dermascan measured dermal thickness and intensity had similarly high ICC values, the %CV values for Dermal intensity are higher compared to that for Dermal thickness for both scar (16.4% versus 6.79%) and normal skin (15.9% versus 5.13%).

The ICC agreement estimates were again high for the thickness and intensity measures, 0.926 (95% CI: 0.898–0.955) and 0.873 (95% CI: 0.826–0.921) respectively. Again the estimates for normal skin were similar.

3.3.4. *Cutometer: R0 and R2 measurements*

Cutometer measurements were only available for 54 patients as 1 patient did not have enough time during the session for the Cutometer measurement. The inter- and intra-rater ICC values for the Cutometer R0 and R2 measures are shown in Table 8.

The R0 measure had acceptable inter- (ICC [Single, Average] =0.715, 0.883) and intra-rater ICC values (ICC [Single, Average] =0.80, 0.92) for scar tissue. The R0 intra-rater ICC values for normal skin was also acceptable (ICC [Single, Average]=0.85, 0.94) but only the ICC (Average) of the R0 inter-rater ICC was above the set threshold (ICC Average=0.827) but not the ICC Single (ICC Single=0.615). In contrast to the R0 value, the R2 measure for scar tissue only had acceptable Average Inter-rater and intra-rater ICC (Average) values (ICC Average=0.758, 0.76) but its Single Inter-rater ICC (Single=0.510) and Intra-

Table 6 – Inter-rater and intra-rater ICC values of the erythema and pigmentation measures for the Scanoskin system for scar tissue and normal skin.

Measure	Scar tissue		Normal skin	
	Erythema	Pigmentation	Erythema	Pigmentation
Mean	78.29	81.12	31.65	98.42
IQR	53.78–108.76	54.90–98.79	13.68–47.82	81.00–116.53
%CV (patient)	48.2	38.5	69.8	24.4
Inter-rater ICC				
Single ICC (95% CI)	0.969 (0.944–0.984)	0.972 (0.951–0.986)	0.926 (0.871–0.961)	0.957 (0.925–0.978)
Average ICC (95% CI)	0.989 (0.981–0.994)	0.991 (0.983–0.995)	0.974 (0.953–0.987)	0.985 (0.974–0.992)
%CV (SEM)	8.55 (3.32)	6.34 (2.59)	18.05 (2.78)	4.42 (2.43)
Intra-rater ICC				
Single ICC (95% CI)	0.995 (0.985–0.998)	0.989 (0.955–0.998)	0.985 (0.962–0.994)	0.994 (0.940–0.997)
Average ICC (95% CI)	0.998 (0.995–0.999)	0.996 (0.995–0.999)	0.995 (0.987–0.998)	0.997 (0.979–0.999)
%CV (SEM)	3.44 (3.31)	3.25 (2.59)	10.4 (2.76)	2.34 (2.43)
Agreement and consistency				
ICC agreement (95% CI)	0.966 (0.945–0.986)	0.965 (0.946–0.985)	0.917 (0.870–0.963)	0.949 (0.921–0.978)
ICC consistency (95% CI)	0.995 (0.993–0.998)	0.989 (0.983–0.995)	0.984 (0.975–0.993)	0.987 (0.979–0.994)

Table 7 – Inter- and intra-rater reliability values for the Dermascan ultrasound thickness and intensity measurements for scar tissue and normal skin.

Measure	Scar tissue		Normal skin	
	Dermal thickness	Dermal intensity	Dermal thickness	Dermal intensity
Mean	2.65	10.59	1.47	28.60
IQR	1.97–3.25	5.72–14.42	1.10–1.64	18.57–38.73
%CV (patient)	40.1	64.4	33.5	46.5
Inter-rater ICC				
Single ICC (95% CI)	0.957 (0.934–0.973)	0.918 (0.874–0.948)	0.967 (0.949–0.980)	0.863 (0.796–0.912)
Average ICC (95% CI)	0.985 (0.977–0.991)	0.971 (0.954–0.982)	0.989 (0.982–0.993)	0.950 (0.921–0.969)
%CV (SEM)	6.79 (0.10)	16.4 (0.90)	5.13 (0.04)	15.9 (2.42)
Intra-rater ICC				
Single ICC (95% CI)	0.951 (0.871–0.987)	0.928 (0.881–0.976)	0.948 (0.881–0.976)	0.931 (0.855–0.966)
Average ICC (95% CI)	0.983 (0.953–0.996)	0.937 (0.947–0.991)	0.982 (0.954–0.993)	0.976 (0.947–0.989)
%CV (SEM)	7.33 (0.10)	13.68 (0.90)	5.35 (0.04)	12.62 (2.42)
Agreement and consistency				
ICC agreement (95% CI)	0.926 (0.898, 0.955)	0.873 (0.826, 0.921)	0.947 (0.927, 0.968)	0.823 (0.758, 0.889)
ICC consistency (95% CI)	0.951 (0.931, 0.970)	0.923 (0.894, 0.953)	0.947 (0.927, 0.968)	0.924 (0.894, 0.954)

Table 8 – Inter- and intra-rater reliability values for the Cutometer R0 and R2 measures for scar tissue and normal skin.

Measure	Scar tissue		Normal skin	
	R0	R2	R0	R2
Mean	0.61	0.77	1.09	0.82
IQR	0.503–0.740	0.720–0.825	9.08-1.26	0.77–0.88
%CV (patient)	31.3	9.4	20.9	8.2
Inter-rater ICC				
Single ICC (95% CI)	0.715 (0.522–0.834)	0.542 (0.385–0.683)	0.615 (0.373–0.772)	0.510 (0.348–0.660)
Average ICC (95% CI)	0.883 (0.766–0.938)	0.780 (0.653–0.866)	0.827 (0.641–0.910)	0.758 (0.616–0.853)
%CV (SEM)	17.15 (0.06)	7.84 (0.03)	14.74 (0.09)	5.90 (0.03)
Intra-rater ICC				
Single ICC (95% CI)	0.80 (0.529–0.930)	0.58 (0.225–0.807)	0.85 (0.687–0.928)	0.52 (0.206–0.783)
Average ICC (95% CI)	0.92 (0.771–0.976)	0.79 (0.465–0.926)	0.94 (0.868–0.975)	0.76 (0.438–0.916)
%CV (SEM)	13.29 (0.04)	7.89 (0.03)	7.97 (0.05)	7.35 (0.03)
Agreement and consistency				
ICC agreement (95% CI)	0.601 (0.485–0.718)	0.373 (0.246–0.501)	0.544 (0.408–0.679)	0.335 (0.210–0.459)
ICC consistency (95% CI)	0.736 (0.646, 0.826)	0.455 (0.325, 0.585)	0.803 (0.724, 0.882)	0.400 (0.271, 0.529)

rater (Single=0.52) values were all below the acceptable threshold. Both the single inter- (ICC Single=0.510) and intra-rater (ICC Single=0.52) ICC values for the R2 measure for normal skin were below the threshold but the average inter- (ICC Average=0.758) and intra-rater (ICC Average=0.76) ICC values reached acceptable levels.

We further analysed our data by dividing the group into scars with mVSS Pliability scores of <2 and above 2 (Table 9). Whilst the inter-rater R0 ICC values for normal scars (ICC [Single, Average]=0.718, 0.884) were better than the firmer hypertrophic scars (ICC [Single, Average]=0.568, 0.798), the R2 inter-rater ICC values for hypertrophic scars showed a reverse trend and were higher than that for non-hypertrophic scars.

3.3.5. *Patient satisfaction with devices questionnaire*

All of the objective measurement devices had over 90% of "Very good" and "Good" ratings in all three of the patient satisfaction parameters of overall satisfaction, comfort and time to measure. Interestingly, high patient satisfaction was also found for the more invasive skin biopsy procedure though slightly lower compared to the non-invasive tools (average %) in terms of overall satisfaction (87.9% versus 96%), comfort (78.8% versus 95.5%), and time to measure (90.6% versus 95.1%).

3.3.6. *Patient rated scar parameter importance*

Responses were available from 54 of the 55 patients for patient rated scar parameter importance (Table 10).

Table 9 – Differences in R0 and R2 inter-rater ICC values between normal scars (mVSS Pliability scores of 2 or less) and hypertrophic scars (mVSS Pliability values of >2).

Measure	Non-hypertrophic scars (mVSS Pliability score <2)		Hypertrophic scars (mVSS Pliability >2)	
	R0 (n=21)	R2 (n=21)	R0 (n=33)	R2 (n=33)
Single	0.718	0.294	0.568	0.567
ICC (95% CI)	(0.431–0.876)	(0.024–0.582)	(0.341–0.747)	(0.370–0.737)
Average	0.884	0.555	0.798	0.797
ICC (95% CI)	(0.695–0.955)	(0.070–0.807)	(0.608–0.899)	(0.638–0.894)
%CV (SEM)	14.47 (0.06)	6.52 (0.03)	18.87 (0.05)	8.68 (0.04)

Table 10 – Wilcoxon signed-rank test comparison of the scar parameters. The lower the rank, the more important the parameter was to the patient**. Parameters with statistically significant correlations are shaded in grey.**

Parameters (p-values)	Mean rank	Std deviation	Redness	Pigmentation	Surface area	Thickness	Pliability	Pain/itch
Redness	3.70	1.74	n/a	–	–	–	–	–
Pigmentation	4.01	1.70	0.16	n/a	–	–	–	–
Surface area	3.70	1.43	0.956	0.217	n/a	–	–	–
Thickness	3.22	1.43	0.195	0.018	0.049	n/a	–	–
Pliability	3.46	1.73	0.574	0.218	0.429	0.258	n/a	
Pain/itch	3.02	2.05	0.115	0.033	0.134	0.55	0.28	n/a

A statistically significant difference in patient perceived scar parameter importance was found (Friedman test, chi square=11.26, p=0.046). Pain and itch were shown to be the most important scar parameter to patients and pigmentation the least. To examine where differences in ranking actually occurred, separate Wilcoxon signed rank tests on the different combinations of the parameters were run. The Wilcoxon signed-rank test showed that patients perceived scar thickness to be significantly more important than pigmentation (p=0.018) and surface area (p=0.049), whereas pain and itch was perceived to be more important than pigmentation (p=0.033).

Due to the multiple comparisons done, a Bonferroni adjustment was performed, with the adjusted significant p-value calculated to be 0.003 (i.e. 0.05/15 comparisons). After adjustment, there were no statistically significant differences between the parameters in terms of perceived importance.

4. Discussion

Accurate and reliable measurements of hypertrophic scarring in burns are increasingly being recognised as important in both clinical practice and research. This is especially true in the field of wound healing where long term scarring is an important outcome measure to determine the effectiveness of a surgical procedure or treatment. In this study, both the intra- and inter-rater reliability of the objective tools, i.e. the Cutometer, the Dermascan ultrasound, the DSM II Colormeter, the Scanoskin system were tested on both scar tissue as well as the matching normal skin sites. Additionally, the inter-rater reliability of commonly used subjective scales, the modified VSS and POSAS were also tested. Comparisons of the objective scar measurement tools with the mVSS, POSAS and also histological parameters to test the validity of the objective measures was performed but will be addressed in a separate study.

This study differs from previous similar studies in a number of different ways. Firstly a power calculation was performed to calculate the minimum number of participants required, and as a result of this, a much larger number of patients have been recruited (n=55) compared to other similar studies which typically recruit around 30 subjects or less, for e.g. Nedelec et al. [10,11], Gankande et al. [23] and Draaijers et al. [24].

Additionally, other reliability studies have performed their measurements over an average 2 week period and not in a single session as in this study [10,11,25]. This predisposes these studies to errors due to relocating the exact site of measurement despite rigorous protocols being utilised and this is a common source of error that is acknowledged by several studies. Thus the ICC values of the tools in these studies are influenced substantially by the reliability of their relocation protocol. However it has to be acknowledged that although testing in a single day reduces error likely associated with relocation, it does not mirror clinical practice and thus actual reliability values could reasonably be expected to be lower at other time intervals.

Our study has shown that the traditionally and widely used subjective scar scales, the modified VSS and POSAS have poor reliability when performed by less than three assessors for the majority of the subscales with the exception of the POSAS vascularity subscale which has an acceptable inter-rater average ICC of 0.78 when performed by 2 assessors. Even with three assessors, the mVSS total score as well as the POSAS pigmentation, relief, pliability, change in surface area and overall score still have lower than acceptable ICC values. The reliability of the POSAS subscales is likely to be even lower as these ICC values were calculated without taking into account the additional categorical classifications each subscale possesses, for example the pigmentation subscale can be rated into the categories of "Hypo", "Hyper" and "Mix" in addition to the 1–10 scale. The requirement of having three assessors for acceptable reliability makes these subjective scores impractical for most clinical and research settings. One explanation for the poor reliability of these subjective scores is the lack of a standardised reference on which to score against. The assessors are thus affected by both recall bias and are also biased by their own personal experience of previously seen scars when judging the assessed scar. Although an increased number of scar assessments may improve the intra-rater reliability of an assessor due to improved familiarity with the scoring scale and the exposure to a wider range of scar severity and types, there is no evidence that it makes the assessment more accurate or improves inter-rater reliability unless feedback is given to the assessor (e.g. comparing own measures with a standard, to others or to an objective measure).

The DSM II a* and narrow band Erythema had acceptable inter- and intra-rater ICC values of above 0.70 for both single and average measures for scar tissue. The inter-rater ICC values for normal skin however were found to be lower compared to that for scar tissue. The results point towards a difference in the readings between raters, most likely due to different pressures being applied to the measured site resulting in varying levels of blanching, though a greater difference would be expected to be seen in scars compared to normal skin. Another explanation could be that the measured normal skin site had been irritated by multiple measurements (and thus appearing more erythematous,) by the other tools and insufficient time was allowed for the skin to recover fully despite time being allocated in the protocol for this. Our results are in contrast with other studies which show that the erythema ICC values for normal skin are usually similar or higher than that for scar sites [10,12].

In terms of pigmentation measurement, the DSM II L*, and narrow band Melanin parameters both show high inter- and intra-rater ICC values of above 0.90 for both single and average measures although the b* measure had poor reliability values which is in contrast with the reliability of the b* value of similar colour measurement systems such as the Labscan XE184 [25]. Our analyses have shown that this is likely to be due to the fact that the b* parameter is affected by both pigmentation ("yellowness") as well as skin circulation (likely venous circulation, "blueness") as evidenced by the significant correlations to both histologically measured melanin concentrations and CD31 measured vessel concentration thus leading to greater variability (data not shown). Compared to the pigmentation measures (other than the b* measure), the objective measurements for erythema of the DSM II Colormeter have lower ICC values. This again is likely due to the erythema parameters being susceptible to the varying amounts of pressure applied onto the area of measurement which may cause different degrees of blanching in contrast to pigmentation parameters which are not influenced by pressure.

The Scanoskin camera system parameters for both erythema and pigmentation have been shown to have greater reliability compared to the DSM II parameters, although more observations are required to confirm this. An additional advantage of the Scanoskin system is that it allows the measurement of the parameters over a much greater area (up to approximately 3500 cm^2, although light exposure at peripheries will be lower than centre,) will be compared to the DSM II which only has a measurement area of 4 mm.

The Dermascan ultrasound system has been shown to have high (ICC > 0.90) inter- and intra-rater ICC (Single and Average) values for both dermal thickness and intensity in scar tissue and normal skin. In addition to this, the dermal thickness parameter also had significant correlations with the subjective height subscale of both the mVSS and POSAS-Observer as well as histologically measure dermal thickness ($p < 0.01$). This is in agreement with most studies on the Dermascan system [10,11,14]. A study by Agabalyan et al. [26] however reported a weak correlation between the Dermascan measured ultrasound and histological measured dermal thickness (Spearman rank correlation of −0.6242). There are some important methodological differences between this study and ours. This includes the small number of samples (n = 10) compared to our study (n = 38), and the use of the H&E stain which will not allow the authors to differentiate between scar tissue and uninjured dermis. A flaw of the Dermascan system however is the inability to adjust the image enhancement (mode) after the image is taken compared to newer systems such as the DUB Skinscanner system (Taberna Pro Medicum, Germany) [27], which makes the detection of the lower dermal border difficult particularly in thicker scars when using the same enhancement mode as the thinner scars.

For pliability measures, the Cutometer R0 parameter had acceptable inter and intra-rater ICC values for scar tissue but the R2 parameter did not reach the required threshold. Previous studies have shown both low and high reliability values for the Cutometer dependent on the parameter being evaluated [10,11,23,24]. In studies that have shown low inter-rater reliability values, this was thought to be due to scars with high rigidity [10,11] however in the study by Nedelec et al. [11] this was found to be only true for the R0 value where the normal skin R0 ICC value was higher than the scar tissue R0 ICC value. The R2 value for normal skin (ICC = 0.55) was however lower than the R2 value for severe scar tissue (ICC = 0.71) in that study which agrees with our study findings. In our study, this difference in reliability between softer and firmer scars had been shown to be true for the R0 parameter but not the R2 parameter (Table 9).

Previous studies have shown both low and high reliability values for the Cutometer are dependent on the parameter being evaluated [10,11,23,24]. In studies that have shown low inter-rater reliability values, this was thought to be due to scars with high rigidity [10,11] however in the study by Nedelec et al. [11] this was found to be only true for the R0 value where the normal skin R0 ICC value was higher than the scar tissue R0 ICC value. The R2 value for normal skin (ICC = 0.55) was however lower than the R2 value for severe scar tissue (ICC = 0.71) in that study which agrees with our study findings.

The study by Nedelec et al also attributed the reduced reliability of the Cutometer to the sensitivity of the device to slight differences in location [10]. However in our study, we repeated the measurements in a single session (which should significantly reduce relocation errors,) in contrast to the Nedelec study where repeat measurements were done over three sessions within a 2 week period, and despite this the reliability of the Cutometer remained lower compared to the other objective devices.

Other studies give further insight into the reasons that may contribute to the lower reliability of the Cutometer. Bonaparte et al showed that removing the Cutometer probe in between measurements resulted in a higher reliability compared to leaving the probe in place [28]. This effect was thought to be due to mechanical interference by the probe which prevents skin recoil if the probe was not removed in between readings. In a separate study by Bonaparte et al., they found that moderate to heavy additional forces (50–500 grams) significantly altered measurements [29]. Muller et al. also showed similar findings that showed different contact forces between the probe and skin surface can cause differences in the initial deformation of the skin which can then lead to differences in measurements [30].

Patients' views and preferences are an important component of developing a new investigation as poor patient acceptance will inevitably lead to poor uptake or compliance. Our study has shown that all the objective measurement devices rated highly with the patients including time to measure and comfort. Anecdotally, many of the patients also found it fascinating to be able to have an alternate way to visualise and monitor their scars on the measurement devices e.g. the cross section ultrasound views and the result curves on the Cutometer program. Thus time and cost permitting, incorporating objective devices (e.g. at a few time intervals in the scar treatment process,) this may further enhance patient engagement in research or therapy.

When looking at the importance placed by patients on the different parameters, although not statistically significant after adjustment, the results may imply that physical comfort (absence of pain and itch) and functionality were more important factors to patients compared to appearance (surface area and colour). However this may be due to the majority of the scar being assessed being >12 months old which means that the scars are more likely to be mature scars with reduced erythema. Additionally, the original findings may be upheld with larger patient numbers thus further investigation is warranted.

This study has shown that the subjective scales mVSS and POSAS have much poorer reliability compared to the objective devices. However subjective scales still have an important role in clinical practice and certain advantages over objective measurement devices. Subjective scales are advantageous over objective devices in terms of cost and availability. Most, if not all, of the subjective scales are free to use and are easily obtainable from the internet whereas the objective devices used in the study range from £5000 to £18000. This is especially important in smaller units and in developing countries where no additional funding for these devices would be available. Subjective scales such as POSAS also allows the patient to rate their own scars and therefore involve the patient in their own care. There are also many limitations in objective devices that are easier to overcome with subjective measures such as measurement in difficult areas such as tight spaces which do not fit the measurement probe or highly mobile areas such as joints and large scars as most of the objective devices can only measure small areas. The limited time available to spend on patients in clinics also means that subjective measures are more easily completed. Although subjective assessment is extremely important as they may include patients' perception, the poor reliability of subjective assessments however mean that they cannot be used solely as an assessment for of scars e.g. in studies looking at the effectiveness of scar treatments. They must be combined with, ideally, a panel of objective scar measures.

There are a number of limitations to this study. Due to the study design, patients only participated in a single session for all measurements, and thus we were not able to calculate the intra-rater reliability for the subjective scar scales (mVSS and POSAS) or any longitudinal data for the scars. A likely contributor to the high reliability of both the Scanoskin and Dermascan measurements is the use of a single researcher to perform all the processing and measurements of the obtained images. This was unfortunately not avoidable due to the time constraints and other work commitments of the raters. For the Scanoskin camera and the majority of the Dermascan images, the use of a single researcher for all measurements is unlikely to have introduced significant bias as the areas that had to be delineated manually for measurement were very clear. However in certain Dermascan images, especially the thicker scars, the inner layer of the dermis can sometimes be hard to visualise and thus some subjectivity is introduced. Although all the clinicians were trained and competent in the use of the devices before the start of the study, at the final analysis it was found that some clinicians had consistently lower reliability in certain devices. This shows there is a need to check the intra-rater reliability during either clinical practice or studies to ensure that problems that lead to lower reliability can be identified and ameliorated.

A high reliability of a measurement does not equate to a high validity and this will be further investigated in a separate related study where the measurements are validated against the mVSS as well as histological markers which are the best available "gold standard" currently. Histological markers however themselves are subject to variability and artefacts due to processing and interpretation and sampling bias.

In summary, we found that the reliability of the subjective scores mVSS and POSAS both fell below the acceptable threshold. For the objective tools, the following had good to excellent intra and inter-rater reliability scores: (1)The DSM II L* and narrow band Melanin measures for pigmentation and the a* and narrow band Erythema values for erythema, (2) The Dermascan thickness and intensity measurements, the (3) Cutometer R0 and (4) The Scanoskin Erythema and pigmentation measures. Therefore these devices are recommended to be used in both clinical practice and research in the measurements of burn scars in combination with patient reported outcome measures such as the Brisbane Burn Scar Impact Profile (BBSIP) [31] to provide a well-rounded evaluation of burn scar and its impact on the patients' quality of life. At present, the high cost of the objective devices, inexperience of clinicians with their use and limited clinical time are major obstacles to their wider adoption and means that the objective devices are largely limited to commercial and research use but it is hoped that this study will promote their use and also spur more research into improving the objective measurement tools and further innovations in objective measures.

Funding

This work was supported by The Scar Free Foundation (Grant code: DLAD.RDAR16265) and the Wellcome Trust (Grant code: HICF-R8-450).

Declarations of interest

None.

REFERENCES

[1] Chipp E, Charles L, Thomas C, Whiting K, Moiemen N, Wilson Y. A prospective study of time to healing and hypertrophic scarring in paediatric burns: every day counts. Burns Trauma 20175:.

[2] Gangemi EN, Gregori D, Berchialla P, Zingarelli E, Cairo M, Bollero D, et al. Epidemiology and risk factors for pathologic scarring after burn wounds. Arch Facial Plast Surg 200810:.

[3] Thompson CM, Hocking AM, Honari S, Muffley LA, Ga M, Gibran NS. Genetic risk factors for hypertrophic scar development. J Burn Care Res 201334:.

[4] Bae SH, Bae YC. Analysis of frequency of use of different scar assessment scales based on the scar condition and treatment method. Arch Plast Surg 2014;41(2):111–5.

[5] Van Der Wal M, Tuinebreijer W, Bloemen M, Verhaegen P, Middelkoop E, Van Zuijlen P. Rasch-analysis of the Patient and Observer Scar Assessment Scale (POSAS) in burn scars. Wound Rep Regen 2010;18(6):A97.

[6] Brown BC, Moss TP, McGrouther DA, Bayat A. Skin scar preconceptions must be challenged: importance of self-perception in skin scarring. J Plast Reconstr Aesthet Surg 2010;63(6):1022–9.

[7] Zhang J, Miller CJ, O'Malley V, Bowman EB, Etzkorn JR, Shin TM, et al. Patient and physician assessment of surgical scars: a systematic review. JAMA Facial Plast Surg 2018;20(4):314–23.

[8] Streiner DL, GR N. Health measurement scales: a practical guide to their development and use. Oxford, NY: Oxford University Press; 2003.

[9] Lee KC, Dretzke J, Grover L, Logan A, Moiemen N. A systematic review of objective burn scar measurements. Burns Trauma 2016;4:14.

[10] Nedelec B, Correa JA, Rachelska G, Armour A, LaSalle L. Quantitative measurement of hypertrophic scar: interrater reliability and concurrent validity. J Burn Care Res 2008;29 (3):501–11.

[11] Nedelec B, Correa JA, Rachelska G, Armour A, LaSalle L. Quantitative measurement of hypertrophic scar: intrarater reliability, sensitivity, and specificity. J Burn Care Res 2008;29 (3):489–500.

[12] van der Wal M, Bloemen M, Verhaegen P, Tuinebreijer W, de Vet H, van Zuijlen P, et al. Objective color measurements: clinimetric performance of three devices on normal skin and scar tissue. J Burn Care Res 2013;34(3):e187–94.

[13] Tan A, Pedrini FA, Oni G, Frew Q, Philp B, Barnes D, et al. Spectrophotometric intracutaneous analysis for the assessment of burn wounds: a service evaluation of its clinical application in 50 burn wounds. Burns 2017;43(3):549–54.

[14] Van den Kerckhove E, Staes F, Flour M, Stappaerts K, Boeckx W. Reproducibility of repeated measurements on post-burn scars with Dermascan C. Skin Res Technol 2003;9(1):81–4.

[15] Myers SL, Cohen JS, Sheets PW, Bies JR. B-mode ultrasound evaluation of skin thickness in progressive systemic sclerosis. J Rheumatol 1986;13(3):577–80.

[16] Fong SS, Hung LK, Cheng JC. The cutometer and ultrasonography in the assessment of postburn hypertrophic scar—a preliminary study. Burns 1997;23(Suppl. 1):S12–8.

[17] Shoukri MM, Asyali MH, Donner A. Sample size requirements for the design of reliability study: review and new results. Stat Methods Med Res 2004;13(4):251–71.

[18] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86(2):420–8.

[19] de Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine: a practical guide. Cambridge: Cambridge University Press; 2011.

[20] Nichols D. Stats Fleiss kappa. 4th ed. SPSS, IBM; 2015 July 24.

[21] McHugh ML. Interrater reliability: the kappa statistic. Biochem Med 2012;22(3):276–82.

[22] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159–74.

[23] Gankande TU, Duke JM, Danielsen PL, DeJong HM, Wood FM, Wallace HJ. Reliability of scar assessments performed with an integrated skin testing device—the DermaLab Combo®. Burns 201440:.

[24] Draaijers LJ, Botman YA, Tempelman FR, Kreis RW, Middelkoop E, van Zuijlen PP. Skin elasticity meter or subjective evaluation in scars: a reliability assessment. Burns 2004;30(2):109–14.

[25] Li-Tsang CW, Lau JC, Liu SK. Validation of an objective scar pigmentation measurement by using a spectrocolorimeter. Burns 2003;29(8):779–84.

[26] Agabalyan NA, Su S, Sinha S, Gabriel V. Comparison between high-frequency ultrasonography and histological assessment reveals weak correlation for measurements of scar tissue thickness. Burns 2017;43(3):531–8.

[27] Herbig LE, Kohler L, Eule JC. High resolution imaging of the equine cornea using the DUB®-SkinScanner v3.9. Tierarztliche Praxis Ausgabe G, Grosstiere/Nutztiere 2016;44(6):360–7.

[28] Bonaparte JP, Chung J. The effect of probe placement on inter-trial variability when using the Cutometer MPA 580. J Med Eng Technol 2014;38(2):85–9.

[29] Bonaparte JP, Ellis D, Chung J. The effect of probe to skin contact force on Cutometer MPA 580 measurements. J Med Eng Technol 2013;37(3):208–12.

[30] Müller B, Elrod J, Pensalfini M, Hopf R, Distler O, Schiestl C, et al. A novel ultra-light suction device for mechanical characterization of skin. PLoS One 2018;13(8):e0201440.

[31] Tyack Z, Kimble R, McPhail S, Plaza A, Simons M. Psychometric properties of the Brisbane Burn Scar Impact Profile in adults with burn scars. PLoS One 2017;12(9):e0184452.