In a clinimetric analysis, 3D stereophotogrammetry was found to be reliable and valid for measuring scar volume in clinical research.
Abstract

Introduction: Volume is an important feature in the evaluation of hypertrophic scars and keloids. Three-dimensional (3D) stereophotogrammetry is a noninvasive technique for the measurement of scar volume. This study evaluated the reliability and validity of 3D stereophotogrammetry for measuring scar volume.

Materials and Methods: To evaluate reliability, 51 scars were photographed by two observers. Interobserver reliability was assessed by the intraclass correlation coefficient (ICC), and the measurement error was expressed as limits of agreement (LoA). To assess validity, 60 simulated (clay) scars were measured by 3D stereophotogrammetry and subsequently weighed (gold standard). The correlation of volumes obtained by both measures was calculated by a concordance correlation coefficient (CCC), and the measurement error was expressed as a 95% prediction interval.

Results: The ICC was 0.99, corresponding to a high correlation of measurements between two observers, although the LoA were relatively wide. The correlation between 3D stereophotogrammetry and the gold standard was also high, with a CCC of 0.97. Again, the plot of the differences and LoA showed moderate agreement for the validity.

Conclusions: Three-dimensional stereophotogrammetry is suitable for the use in clinical research but not for the follow-up of the individual patient.

Introduction

Scarring after burns and after surgical and traumatic injury may lead to functional, esthetic and psychological problems. The development of pathological scars is regularly seen especially after burns. One of the most distinct features of these pathological scars is that they are elevated above the surrounding skin, thereby having an increased volume. An adequate assessment of the scar property volume allows clinicians and researchers to compare the effectiveness of therapy protocols. Scar properties may be assessed by validated scar assessment scales, such as the POSAS. There is, however, a need for objective scar assessment tools, to quantitatively measure the volume of a scar and to provide an objective patient follow-up after applied treatment.

Studies describing techniques that objectively measure scar volume are scarce. A reliable technique is taking a cast of the scar with elastomer putty (negative impression), where after the resulting mold is filled with plaster and subsequently weighed. This casting method though has several disadvantages: it may cause patient discomfort, the plaster density may vary, and it is time consuming. Moreover, the use is limited to small scars surrounded by normal skin. The last few years, several three-dimensional (3D) imaging methods have become available for the assessment of volume in clinical and research settings. One of these methods is 3D stereophotogrammetry, which uses digital color images that are captured simultaneously by two cameras. Especially in the field of oral and maxillofacial surgery, various studies have been performed to identify facial landmarks, to measure volumetric changes in cleft lip and palate nose, and to image facial surface area.

Three-dimensional stereophotogrammetry for measuring scar volume was tested in two studies with an experimental set up and found to have a good validity (both Pearson’s correlation coefficient >0.97). Although useful, the applicability of these results is limited mainly because of the small inclusion numbers. For assessing the appropriateness of a measurement device, both reliability (i.e., the degree to which repeated measurements provide similar results) and validity (i.e., the degree to which the device measures what it is intended to measure) are essential features. The objective of this study was to evaluate the reliability and validity of 3D stereophotogrammetry for measuring scar volume.
Materials and Methods

Study population
Patients with clinically raised scars (hypertrophic scars and keloids) were recruited from the scar outpatient clinics in the Red Cross Hospital in Beverwijk and the VU Medical Center in Amsterdam from August until October 2012. Scars with a diameter of less than 20 cm fitted the measuring frame of the camera and were eligible for inclusion. In addition, healthy volunteers, either working as health care workers in the Burn Center or as intern at the surgery department of the Red Cross Hospital, participated in the validation part of the study. Verbal informed consent was obtained from patients and volunteers. The principles outlined in the declaration of Helsinki were followed.

The imaging system
The stereophotographic system 3D LifeViz (Canon 500D body, Dermapix software, Quantificare, Sophia Antipolis, France) was used. This system includes a customized Canon 500D, 15.1 megapixel digital reflex camera with a 39-mm-bifurcated lens. Photographs were taken from a standardized distance of 60 cm by the use of two light beams that converge at one spot when the camera is held perpendicular at the correct distance from the targeted point (Figure 1a). The camera captures two images simultaneously by taking a photograph from different angles (Figure 1b). The software program Dermapix automatically integrates the stereo images and produces a three-dimensional reconstruction in the ‘fine analysis’ mode. The margin of the scar was manually marked in the software program Dermapix, using a computer mouse. To do so, the image can be pictured from above in a 2D manner. After the margin was defined, the 3D reconstruction shows a black surrounding line (Figure 1c). Thereafter, the software program automatically defines the cutoff plane at the bottom of the scar (i.e., the closing surface), using the curvature of the surrounding skin as a reference. It is possible to adjust the sigma, which defines to what extent ‘the closing surface’ follows this curvature. In this study the sigma was set to 1 to perform volume measurements. Finally, the software program automatically calculates the scar volume by height x width x depth dimensions displayed in a grid (Figure 1c).

Gold standard
To obtain a gold standard, scars with a random form and random volume were created with self-hardening modeling clay (DAS Terracotta, modeling material, Fila Group, Italy). The specific gravity of the clay (\( p = 1.66 \, \text{g/cm}^3 \); range 1.63-1.71 \, \text{g/cm}^3 ) was calculated by averaging five repeated gravity measurements, performed under standard laboratory conditions. The scar models were weighed with a high precision scale (Sartorius MC1 Analytic AC 120S) to provide the actual volume by the formula: volume (mm\(^3\)) = mass/specific gravity.

Figure 1a. Camera held at a fixed distance of 60 cm from the scar.
Figure 1b. Stereographic images captured from different angles.
Figure 1c. Three-dimensional reconstruction by the use of Dermapix software. The black line indicating the scar contour, marked manually.
Study procedure
To assess the reliability, the scars of the patients were photographed by two observers. Both observers were experienced in taking the photographs and using the software to calculate the scar volume. Each observer independently performed a volume measurement of the photographs made by the other observer, resulting in two volume measurements per scar. This procedure was expected to simulate the routine clinical practice best: someone else than the photographer may perform volume analysis. To establish the reliability, both 3D measurements were compared.

The validity was tested by applying the scar models to three different body sites of the volunteers (thorax, upper leg, and lower arm), which represented a flat, moderate, and strong curvature of the body, respectively. Subsequently, two observers took one photograph of the simulated scars, and volume measurements were performed crosswise.

Statistical analysis
Data were analyzed using SPSS 18.0 (SPSS Inc., Chicago, USA) and Mplus 6.1 (Muthén & Muthén, Los Angeles, California, USA). General patient and volunteer characteristics were documented. For the statistical analysis, a similar approach was used as described in a previous study by Stekelenburg et al where the device was tested for measuring scar surface area. We refer to the appendix of this study for a detailed description of the statistical calculations. All analyses were performed with the scars as unit of analysis. Because of the skewed distribution of the measurements, data were transformed using a cube root transformation to approximate a normal distribution and constant error variance best. The interobserver reliability was defined as the correlation between the measurements of the two observers and expressed through the interobserver intraclass correlation coefficient (ICCinter). The ICC inter was calculated using the estimated variance components of a scars × observers random-effects model without replications (on the transformed data): scar variance (σ²scar), observer variance (σ²obs), and the error variance (σ²error). The ICC is the ratio between the scar variance and the total variance:

\[
\text{ICC} = \frac{\text{σ}^2_{\text{scar}}}{\text{σ}^2_{\text{scar}} + \text{σ}^2_{\text{obs}} + \text{σ}^2_{\text{error}}}.
\]

The standard error of measurement (SEM) was calculated as:

\[
\text{SEM} = \sqrt{\text{σ}^2_{\text{obs}} + \text{σ}^2_{\text{error}}}
\]

Assessment of absolute agreement between observers, Bland and Altman plots were obtained in which limits of agreement were indicated. These plots show the mean of measurements of two observers on the x-axis against the difference between the two measurements on the y-axis, accompanied by the Limits of Agreement (LoA) and calculated as:

\[
\text{LoA} = \bar{X} \pm 1.96 \times \text{SEM}
\]

In order to assess the validity, the measurement data of 3D stereophotogrammetry and of the gold standard were compared by using the concordance correlation coefficient (CCC). For each of the two measurements, the CCC was calculated, and these two CCC’s were averaged. To quantify the difference between both methods expressed in mm³, again LoA were calculated. In this way, a (nonlinear) regression line and a corresponding 95% prediction interval for the gold standard based on one 3D measurement was calculated using the method of inverse prediction.

Results

Reliability
Fifty-one clinically raised scars of 31 patients were included. Patient demographics and scar characteristics are shown in Table 1. The mean volume, obtained by 3D measurement, was 1619 mm³ (standard deviation (SD): 4120 mm³). After cube root transformation of the data, the ICCinter was found to be 0.99, corresponding to the correlation between the cross 3D volume measurements of two photographs produced by two observers. The variance components were estimated at 25.35 (scars), 0.00 (observers) and 0.10 (error). The SEM was calculated at 0.32. Acquired from the reasonably high SEM, also the LoA are quite wide. Transformation to the traditional Bland-Altman format yields Figure 2, which represents the volumes up to 3000 mm³.

In order to assess the validity, the measurement data of 3D stereophotogrammetry and of the gold standard were compared by using the concordance correlation coefficient (CCC). For each of the two measurements, the CCC was calculated, and these two CCC’s were averaged. To quantify the difference between both methods expressed in mm³, again LoA were calculated. In this way, a (nonlinear) regression line and a corresponding 95% prediction interval for the gold standard based on one 3D measurement was calculated using the method of inverse prediction.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Value</th>
<th>Characteristics</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scars, n</td>
<td>51</td>
<td>Cause of scar, n</td>
<td>11 (22%)</td>
</tr>
<tr>
<td>Patients, n</td>
<td>31</td>
<td>Burn injury</td>
<td>2 (4%)</td>
</tr>
<tr>
<td>Gender, n</td>
<td></td>
<td>Trauma</td>
<td>13 (25%)</td>
</tr>
<tr>
<td>Male</td>
<td>16</td>
<td>Operation</td>
<td>6 (12%)</td>
</tr>
<tr>
<td>Female</td>
<td>15</td>
<td>Injection</td>
<td>13 (25%)</td>
</tr>
<tr>
<td>Age, years</td>
<td>33 (16)</td>
<td>Cause unknown</td>
<td>6 (12%)</td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>2</td>
<td>Scar location, n</td>
<td>7 (14%)</td>
</tr>
<tr>
<td>Maximum</td>
<td>73</td>
<td>Face</td>
<td>12 (23%)</td>
</tr>
<tr>
<td>Scar type, n</td>
<td></td>
<td>Presternal</td>
<td>8 (16%)</td>
</tr>
<tr>
<td>Hypertrophic</td>
<td>15 (29%)</td>
<td>Deltoid region</td>
<td>14 (27%)</td>
</tr>
<tr>
<td>Keloid</td>
<td>35 (69%)</td>
<td>Extremities</td>
<td>6 (12%)</td>
</tr>
<tr>
<td>Mixed</td>
<td>1 (2%)</td>
<td>Back/scapula</td>
<td>4 (8%)</td>
</tr>
<tr>
<td>Ear/lobe</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Patient characteristics.
Validity
Sixty scar models were applied to 20 healthy volunteers. The mean scar volume by weighing the clay and calculating the volume was 5550 mm³ (SD: 4042 mm³). The mean volumes and standard deviations measured by two observers using 3D stereophotogrammetry were, respectively, 5459 mm³ (SD 4256 mm³) and 5629 mm³ (SD 4322 mm³). The mean CCC was 0.97, corresponding to the correlation between 3D stereophotogrammetry measurement and by weighing the clay. The data of validity are plotted in Figure 3, with the gold standard (G) on the y-axis and 3D stereophotogrammetry values (X) on the x-axis. Inverse prediction yielded the (nonlinear) regression line $G = (0.93 \sqrt{X} + 0.12)^3$ and $(0.93 \sqrt{X} + 0.12 \pm 1.74)^3$ for the accompanying LoA. With the regression line, a gold standard value can be calculated from a given 3D measurement. Also, a 45° line (i.e., the line of equality) is plotted, representing 100% agreement between the two methods for measuring scar volume. Note that the prediction line is positioned below the 45° line and that the LoA are again considerably wide.

Discussion
In this study, 3D stereophotogrammetry was found to have a very high reliability expressed by a high ICC$_{reli}$(0.99). Moreover, the CCC that was found for the correlation between 3D stereophotogrammetry and the gold standard was very high (0.97). The reliability and validity of 3D stereophotogrammetry were excellent from this perspective and correspond to other studies that described the use of 3D stereophotogrammetry in measuring volume of clinically raised skin scars. Both studies calculated and presented a very high correlation coefficient (Pearson’s correlation coefficient >0.97) between volumes obtained with 3D stereophotogrammetry, and the actual volume assessed by weighing and calculating the volume of a scar model. Therefore, it can be concluded that not only our study but also all clinimetric studies on 3D stereophotogrammetry for volume measurements of scars showed an excellent reliability and validity measured by correlation statistics. However, although the correlation coefficient is rather a widely accepted and common parameter to assess the reliability and validity, it is not very informative. ICC as well as a Pearson’s correlation coefficient, are relative measurements of correlation and do not provide direct information about the absolute measurement error. For the clinical follow-up of patients, we are interested in the absolute measurement error of an individual measurement because that determines the change that can be detected beyond the measurement error. The SEM’s we found are too large for use in clinical practice. For research purposes,
(i.e., group comparisons), measurement errors are much smaller as these are leveled out in groups because the error is divided by the square root of the number included in the study\(^4\).

Because we were interested in applications of 3D stereophotogrammetry in both research and clinical practice, we additionally studied the absolute agreement between observers as well as the agreement between stereophotogrammetry and the gold standard. This analysis showed that the agreement between observers, in terms of measurement error and LoA, was moderate, which can be concluded from the relatively broad LoA in the Bland and Altman plot. Also, the true agreement between 3D stereophotogrammetry and the gold standard was moderate, which can be interpreted from the considerable width of the LoA. These findings are similar to those of a study that was previously performed by our research group, where the reliability and validity of 3D stereophotogrammetry for the measurement of scar surface area was studied: a high ICC in combination with a moderately good agreement\(^16\). In most studies on reliability and validity, including those previously mentioned, no additional agreement analysis is done\(^8,12\). Likely, the same discrepancy would be found in these studies if additional agreement analysis was performed. For clinimetric studies, we therefore advocate the use of both ICCs, measurement error, and Bland and Altman plots with LoA to have a better (true) understanding of the clinimetric quality.

The discrepancy between the high correlation values and the moderate agreement can be explained by the heterogeneity of the study population. Reliability refers to the question to what extent the scar volumes can be distinguished from each other by the two observers. It is obvious that it is easier to distinguish scars with a wide range of volumes than scars with equal volumes. The ICC is a parameter that is dependent on the heterogeneity of scar volumes in the study population\(^17\). The measurement error refers to absolute differences between the estimation of the volumes, expressed in mm\(^3\). We chose a study population with a wide range of scar volumes because this matches the patient population that is seen in clinical practice. In cases where a high ICC is possibly due to a heterogenic study population, it is important to also assess agreement parameters\(^16,17\).

Possible explanations for the moderate agreement are multiple. First, it was found that scars on extremely curved body parts could not be captured in a single photograph. From a single position of the camera, the edges of the scar were not seen. The 3D reconstruction that was obtained subsequently showed a deformation of the scar edges, resulting in an aberrant volume measurement. Second, it was seen that in some scars, especially keloids, a volumetrically smaller base could not be captured with one photograph. Because the camera was held perpendicular to the scar and the angle of the lenses was predefined, it was not possible to picture the small base resulting in a deformed 3D reconstruction.

For both situations, we suggest taking multiple pictures of a scar from more than one angle and identifying the different segments of a scar with a marker. This will result in a better quality of the photographs without deformation. Subsequently, the volumes of the segments derived with 3D measurement can be summed. Third, another possible reason for the moderate agreement is the three-dimensional aspect of volume measurements. For measuring scar surfaces, quadratic units are used, and for scar volumes, cubic units. This means that measurement errors may occur in three dimensions instead of one. This hypothesis is supported by the finding that a cube root transformation of the data results in an approximate normal distribution of the volume values\(^16\). Fourth, although we optimally standardized the method of taking the pictures and measuring the volumes, measurement errors may arise when taking a photograph. These measurement errors may arise from differences in distance to the object and the exact angle in which the photograph is taken. Technical adjustments to the device might overcome this source of possible measurement error and could be food for future research. Finally, regarding the moderate validity of the device, this could be partially due to the difficulty in determining the exact cutoff point on the backside of the scar. Although we standardized the cutoff point for each scar, the actual curve of the scar, underneath the skin, remains unclear. This difficulty applies though, for all other types of photographic 3D imaging, but also for the gold standard.

In summary, 3D stereophotogrammetry is suitable for the use as measurement instrument in research settings. This is a valuable finding in the search for objective tools to monitor the effect of different treatments. For the use in clinical practice, where we are interested in the individual follow-up of patients in time, the agreement should be improved.
References


