Validated Assessment Scales for the Mid Face

Jean Carruthers, MD,* Timothy C. Flynn, MD,†‡ Thorin L. Geister, PhD,§ Roman Görtelmeyer, PhD,§ Bhushan Hardas, MD,¶ Silvia Himmrich,MSc,§ Derek Jones, MD,** Martina Kerscher, MD, PhD,†† Maurício de Maio, MD,‡‡ Cornelia Mohrmann, MD,‡ Rhoda S. Narins, MD,§§,¶¶ Rainer Pooth, MD, PhD,§ Berthold Rzany, MD, ScM,*** Gerhard Sattler, MD,††† Larry Buchner,BA,‡‡‡ Ursula Benter, MSc,§§§ Lusine Breitscheidel, MD, MPH,§§§ and Alastair Carruthers, MD, FRCPC¶¶¶

BACKGROUND The improvement of aesthetic treatment options for age-related mid face changes, such as volume loss, and the increase in patient expectations necessitates the development of more-complex and globally accepted assessment tools.

OBJECTIVE To develop three grading scales for objective assessment of the infraorbital hollow and upper and lower cheek fullness and to establish the reliability of these scales for clinical research and practice.

METHODS AND MATERIALS Three 5-point rating scales were developed to assess infraorbital hollow and upper and lower cheek fullness objectively. Twelve experts rated identical mid face photographs of 50 subjects in two separate rating cycles using the mid face scales. Test responses of raters were analyzed to assess intra- and interrater reliability.

RESULTS Interrater reliability was substantial for the infraorbital hollow, upper cheek fullness, and lower cheek fullness scales. Intrarater reliability was high for all three scales. Both of the cheek fullness scales yielded higher reliabilities when three rather than two views were used to assess the volume changes of the cheek.

CONCLUSION The mid face scales are reliable tools for valid and reproducible assessment of age-related mid face changes.

Merz Pharmaceuticals GmbH provided funding for this study. Drs. Alastair Carruthers, Jean Carruthers, Derek Jones, Maurício de Maio, Rhoda S. Narins, Berthold Rzany, Martina Kerscher, Timothy C. Flynn, and Gerhard Sattler are paid consultants to Merz Pharmaceuticals.

As we age, changes occur in the mid face. Remodeling of underlying cartilaginous and bony elements combined with soft tissue changes and photodamage causes loss of cutaneous elasticity. All of these changes contribute to the appearance of an aging face.[1–4] The loss of the

*Department of Ophthalmology and Visual Sciences, University of British Columbia, Vancouver, British Columbia, Canada; †Department of Dermatology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina; ‡Cary Skin Center, Cary, North Carolina; §Research and Development, HQ MERZ Pharmaceuticals GmbH, Frankfurt, Germany; ¶Research and Development, MERZ Pharmaceuticals LLC, Greensboro, North Carolina; **Department of Dermatology, University of California at Los Angeles, Los Angeles, California; ††Division of Cosmetic Sciences, University Hamburg, Hamburg, Germany; ‡‡Clinica Mauricio De Maio, Sao Paolo, SP, Brasil; §§Dermatology Surgery and Laser Center, New York, New York; ¶¶Department of Dermatology, New York University School of Medicine, New York, New York; ***Division of Evidence-Based Medicine, Klinik für Dermatologie, Charité – Universitätsmedizin Berlin, Berlin, Germany; †††Rosenparkklinik, Darmstadt, Germany; ‡‡‡Canfield Scientific Inc, Fairfield, New Jersey; §§§INC Research GmbH, Munich, Germany
¶¶¶Department of Dermatology and Skin Sciences, University of British Columbia, Vancouver, British Columbia, Canada

© 2012 by the American Society for Dermatologic Surgery, Inc. • Published by Wiley Periodicals, Inc. •
ISSN: 1076-0512 • Dermatol Surg 2012;38:320–332 • DOI: 10.1111/j.1524-4725.2011.02251.x

suborbicularis oculi fat pads results in the appearance of a tear trough or infraorbital hollow. The infraorbital hollow (tired eyes) can also be attributed to additional factors such as sleep deprivation, stress, extreme diets, and genetic make-up.[5]

Aging effects seen in the cheek area usually result from the volume of the mid face (malar and buccal fat pads) moving downward. As a result, the heart-shaped face of youth, a Caucasian facial aesthetic, changes to the squared facial triangle with the apex at the chin typical of the older face.[5]

Fibrous structures such as the orbicularis retaining ligament (malar septum) and the superior and lateral cheek septa for the mid face compartmentalize facial fat.[6,7] Perforator vessels running in the septal boundaries of each subcutaneous fat compartment ensure the blood supply,[8] but these changes in facial contours are not always due to aging alone, but can also be the result of disease-specific processes, for example the facial lipoatrophy due to the human immunodeficiency virus (HIV) that gives rise to marked upper, mid, and lower facial hollowing and wrinkling of the skin in the mid face.[9] Facial lipoatrophy in people with HIV may progress toward nearly complete subdermal facial fat loss. Contrary to HIV-associated lipoatrophy, minimal fat loss in various facial regions accompanied by generalized facial tissue ptosis often characterizes facial lipoatrophy in aging people without HIV.[10] The mid face regions addressed in this publication are described in more detail in Figure 1.

The increase in treatment options available enables tailoring of treatments to individual needs and goals. There has been a dramatic rise in the number of treatments administered in the past decade.[11] Changes in patient demographics, for example the growing number of men and members of diverse ethnic groups seeking aesthetic facial treatment, have contributed to the current renaissance.[2]

There has been a shift in recent years from a two-dimensional focus on the aesthetic treatment of superficial facial rhytides to an increased understanding of the three-dimensional aspects, which are of particular and striking importance in the aging of the mid face (the loss of volume).[2] There is also a clear trend away from overcorrection, including the "too pulled" look of repeated facelifts, toward a softer, more natural look.

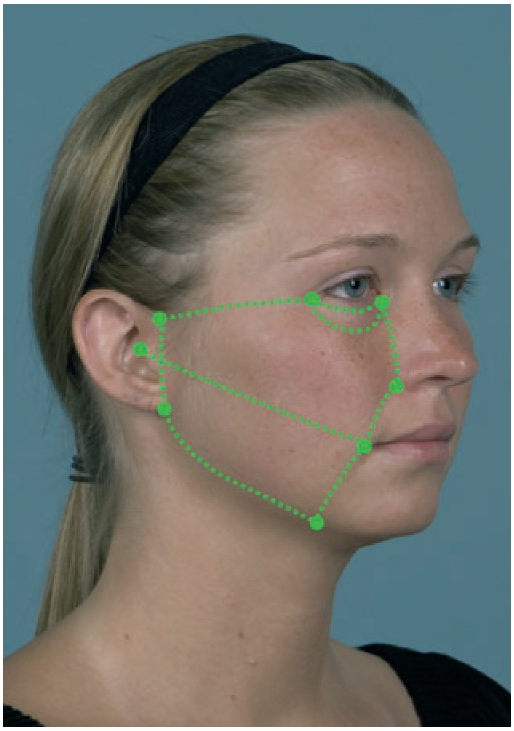

Figure 1. Definition of mid face regions. The upper cheek is bounded superiorly by the infraorbital hollow and its extension to the junction of the attachment of the superior helix of the ear to the face. The medial boundary is the lateral wall of the nose and the nasolabial fold extending to the lateral commissure of the mouth. The inferior boundary of the upper cheek extends from the lateral commissure of the mouth to the superior border of the tragus. This line serves also as the superior boundary of the lower cheek. The medial boundary of the lower cheek is the melomental fold, and the inferior boundary extends along the jawline posteriorly to the lobule of the ear.

The key to rejuvenating the midfacial area is volume restoration through reflation and recontouring.[2] Traditional dermal fillers for facial recontouring include bovine collagen, silicone, and autologous fat,[11,12] but as the use of these early agents has declined and as more patients do not want surgery, biodegradable dermal fillers and volumizers such as cross-linked hyaluronic acid have become the predominant nonsurgical treatment option to correct infraorbital hollows and restore cheek fullness.[5,13] Hyaluronic acid lasts longer and elicits a significantly less frequent allergic response than most collagen products. In addition, hyaluronic acid does not require skin tests, and any posttreatment irregularity can easily be shaped using local massage or application of hyaluronidase.[2,11,13] Calcium hydroxyl-apatite or poly-L-lactic acid (PLLA) is also frequently injected for volume enhancement.[11]

The rapid transformation taking place in global aesthetic medicine calls for the use of more-refined and multifaceted methods to assess facial features objectively r to meet the demanding expectations of patients. Self-explanatory application and easy reliability for long-term patient follow-up are additional attributes that a standard measure should possess. Numerous facial instruments are currently in use.[14] There are no validated, reliable, sensitive tools that are globally accepted in aesthetic clinical research and daily practice that could serve as common foundation for measuring and communicating research and treatment results.[14] MEDLINE searches using the non-Medical Subject Heading terms (*infraorbital hollow* OR *cheek volume* OR *cheek fullness*) AND *validated scale* did not yield any relevant publications.

This article describes three visual mid face scales based on simulated photonumerical image changes for the assessment of the infraorbital hollow and upper and lower cheek fullness. They have been developed to provide an innovative means of measuring these discrete human mid face areas in an objective and reproducible way. The new aesthetic scales are intended to provide a valuable tool for assessing the success of aesthetic mid face treatments in the context of clinical trials or daily practice in aesthetic dermatology and cosmetic surgery. The mid face scales are intended to be used independently to focus on single areas or in combination with additional newly developed assessment scales for other aesthetic units for a more intricate facial evaluation.[15]

Two of the three new mid face scales assess the age-related volume changes of the cheek. Interrater reliabilities obtained at the Scale Summit II for these two cheek fullness scales were only fair or moderate. These validation results thus did not satisfy the quality standards of the team for this important area. The cheek fullness scales were therefore improved to take the three-dimensionality of volume changes of the cheek into better account through the addition of multiple views. These updated scales were then re-assessed in a new validation. The methodology and the reliability results for the original and updated scales are described below.

Methods

Subject Selection

For the collection of the photographs used in the scale creation and validation process of the original scales, 359 men and women aged 25 to 66 with Fitzpatrick skin types I through IV were recruited and screened at three sites in 2009. For the two updated scales, 148 men and women aged 25 to 71 were recruited at four sites in 2010.

Subjects were not to have facial hair, scarring, any previous treatment with toxins or fillers, any surgical procedures in the area of interest (treatment naïve), or HIV-related lipoatrophy. The chosen population was within the framework of the Scale Summit I[16–21] and represents a spectrum that is neither too diverse nor too specific, as in a clinical study. Each screened subject was informed about the objectives and targets of this study and consented to the use of the photographs to be rated and analyzed and used for this and other publications for scientific purposes.

Photographic Methods

Photographs obtained for the creation and validation of all facial scales were two- or three-dimensional images of all qualified subjects. Twenty facial scales were validated at the Scale Summit II to allow for global assessment of the face of each subject. Photographs were taken of the whole frontal face at rest, whole frontal face with hyperkinetic forehead lines, whole frontal face with glabellar frown lines at maximum frown and mouth while pursing, whole lateral face at a 45º oblique view smiling, whole lateral face at a 45º oblique view at rest, and whole lateral face at a 90º view at rest. For the creation of an individual scale, only the views of interest were used. Photographs from above were taken for a "chin down" view to update the cheek fullness scales.

All two-dimensional photographs were taken with a high-resolution photography system using a Nikon D3 (12.1 MP) full frame (35 mm) digital SLR camera (Nikon Corporation, Tokyo, Japan), Nikkor 60-mm f2.8 lens, and specifically configured studio strobes. Equipment for three-dimensional stereophotogrammetry included the Canfield VECTRA CR10 Capture system, a tripod-mounted 3D camera system for 180º capture. The system consisted of two optical pods in a clinical facial field 25-mm focal-length configuration (each pod included three cameras: two monochrome and one high resolution color); Canfield IntelliFlash strobes (two standard) and calibration standard; the Canfield 3D Capture Application for capturing, viewing, and exporting three-dimensional object data; and a personal computer with liquid crystal display monitor. Only two-dimensional photographs were used for the scales and the validation booklets. Three-dimensional material of the subjects was used to create new two-dimensional pictures for the updated cheek fullness scales and the required validation pictures. In particular, this refers to the newly introduced chin-down view, which was not captured with two-dimensional photographs.

Creation of the Photonumerical Rating Scales

Each of the three newly developed grading scales for the mid face was created as a new 5-point rating scale based on computerized photonumerical images. For the mid face unit, there were three scales comprising one scale each for the following mid face traits (Figure 1): the infraorbital hollow representing the major form of suborbicular deficiency from medial to lateral and including the superior part of the tear trough as the minor and only medial form of volume deficiency; upper cheek fullness covering suborbicularis oculi fat deficiency, tear trough, and zygomatico-malar area; and the mid face unit representing lower cheek fullness. The scales were developed to assess the infraorbital hollow (including the tear trough), upper cheek fullness (covering suborbicularis oculi fat deficiency and the zygomatico-malar area), and lower cheek fullness. Facial expression for all scales was at rest. The frontal view was used for the infraorbital hollow scale and the original versions of the upper and lower cheek fullness. A 45º oblique view and a view from above (chin down) was introduced for the update of the cheek fullness scales to take the three-dimensionality of volume changes of the cheek into account.

The photographs were cropped to show the focused aesthetic area of interest (for the mid face, the upper cheek area and the lower cheek aesthetic area). For each scale, a base image was selected from the collection of photographs based on the region of interest and image quality and clarity to be used for the image modification method. Additional images from the photographic database were selected to superimpose varying degrees of severity onto the images based on the evident severity of aging processes in the respective area in the collection of subject portraits. The team and an aesthetic dermatologist reviewed and approved the base image, the images selected for the scale creation, and the description of each numerical grade for each image category. The scale grades for each scale were defined to range from 0 (no sign) to 4 (very severe or marked signs). Photographs used as base images were not used again in the scale validation process. The area of interest for the frontal upper cheek fullness scale view was enlarged for

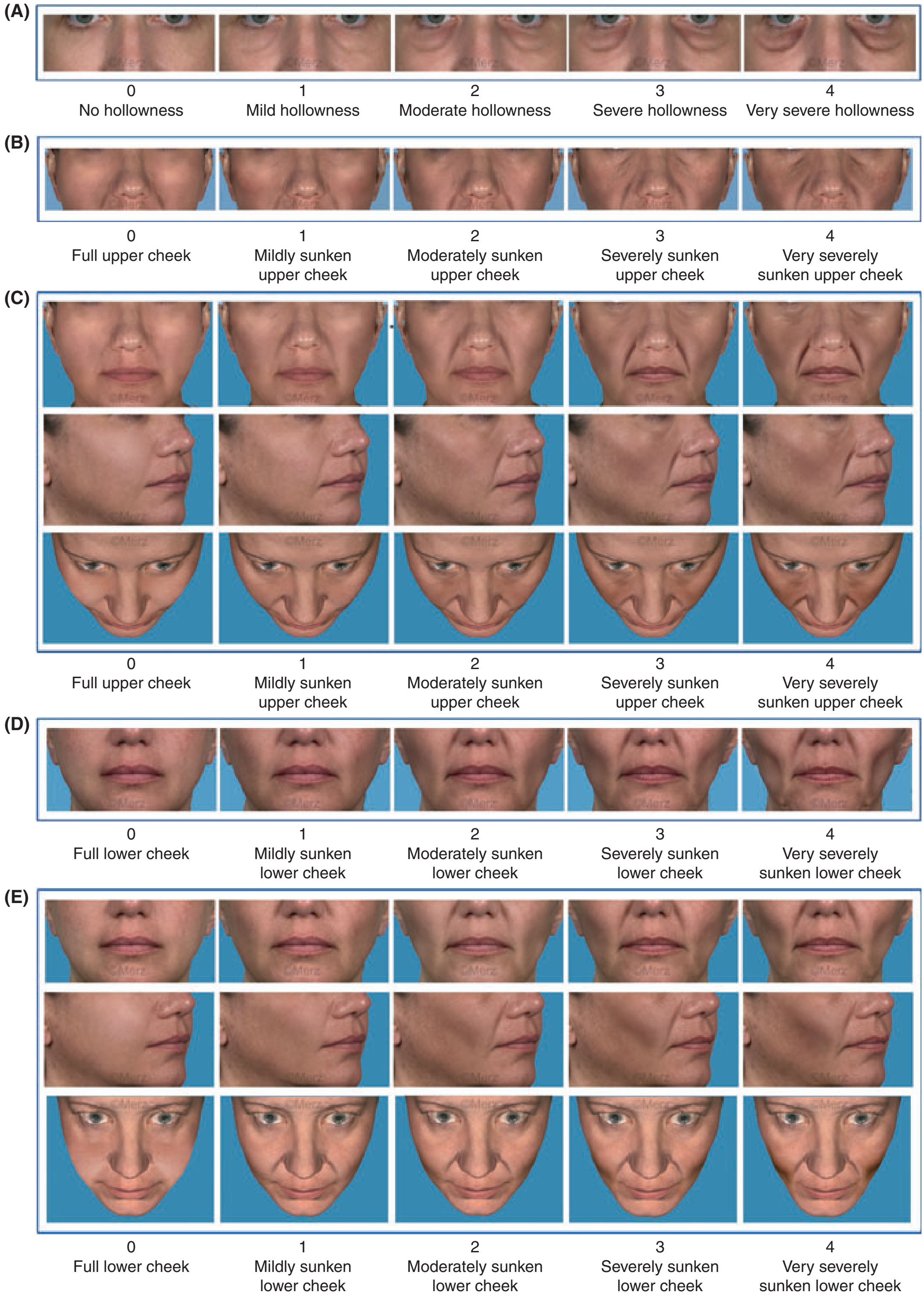
(A)
0
No hollowness
1
Mild hollowness
2
Moderate hollowness
3
Severe hollowness
4
Very severe hollowness
(B)
0
Full upper cheek
1
Mildly sunken upper cheek
2
Moderately sunken upper cheek
3
Severely sunken upper cheek
4
Very severely sunken upper cheek
(C)
0
Full upper cheek
1
Mildly sunken upper cheek
2
Moderately sunken upper cheek
3
Severely sunken upper cheek
4
Very severely sunken upper cheek
(D)
0
Full lower cheek
1
Mildly sunken lower cheek
2
Moderately sunken lower cheek
3
Severely sunken lower cheek
4
Very severely sunken lower cheek
(E)
0
Full lower cheek
1
Mildly sunken lower cheek
2
Moderately sunken lower cheek
3
Severely sunken lower cheek
4
Very severely sunken lower cheek

Figure 2. Mid face aesthetic scales. (A) Infraorbital hollow. (B) Upper cheek fullness. (C) Upper cheek fullness, updated. (D) Lower cheek fullness. (E) Lower cheek fullness, updated.

the update of the cheek fullness scales. The severity spectrum was reduced for the lower cheek fullness scale because the highest grades were considered more severe than normal aging in that area. The subject for the base image stayed the same, and the additional two views were based on that as well. The different angle of the chin-down view was chosen for the upper and lower cheek fullness scales to represent the best possible view of the area of interest from above. See Figure 2 for the final rating scales for the infraorbital hollow and the original and updated scales for the upper and lower cheek fullness.

Validation of the Rating Scales

Validation booklets containing facial images of a subject alongside the appropriate rating scales and fields to enter the ratings were used for the validation. The scale-creation team preselected approximately 100 photographs for validation of the new mid face scales based on quality, content, and equal distribution across each scale. The photographs were based on quality, content, and equal distribution across each severity grade on the scales. An independent clinical expert (board-certified dermatologist) who was not involved in the scale creation and validation process selected50 photographs (25 female, 25 male) for the validation booklets. Thus, the raters who performed the validation were blinded to the overall selection of subjects. The number of subjects for the Scale Summit II (50) was greater than for the Scale Summit I (35),[16–21] which is a common basis for such validations.[22–24] With five severity grades for each scale, each grade could be represented approximately seven times with 50 subjects. Final subjects were chosen based on the severity of their aging-related facial changes so that the severity grades of all scales were represented. Once the photographs had been selected for the validation set, they were cropped to show the respective aesthetic units to be assessed, including the mid face unit. Although three-dimensional photographs were taken of each subject and used for scale creation, only two-dimensional views were included in the validation booklets for the validation of the original mid face scales. Only one individual subject was used for each validation booklet (i.e. all available scales were tested on that subject with the appropriate aesthetic unit of that scale).

The booklets were high-quality double-page printed spiral-bound booklets with unique identifiers (rater name, randomization number). Two validation booklets (landscape letter format) for each chosen subject were produced for each rater to be used for the two validation cycles. A standardized computer randomization program was used to generate unique randomization lists for the sequence of validation booklets of each rater and each validation cycle, resulting in 1,200 individual booklets (12 raters × 50 subjects × 2 validation cycles).

After an introduction of the concept and the evaluation procedure for each aesthetic unit, 12 experts in the field of aesthetic dermatology each rated 50 subjects (50 booklets) presented as cropped photographs alongside the aesthetic scales of the mid face (the 5-point rating scales for infraorbital hollow, upper cheek fullness, and lower cheek fullness) and respective fields to enter the final rating. Each expert independently performed the assessments in two validation cycles, with the second validation cycle taking place within 4 weeks of the first validation cycle. The first validation cycle took place at the Scale Summit II meeting in Berlin, Germany, October 8 and 10, 2009, under standardized conditions. The validation booklets were shipped to each expert for the second validation cycle. The experts were instructed to assess the booklets independently and to return them. The data were entered into a database using the double-entry method and subjected to quality control. Results for the mid face scales are presented here.

For validation of the updated upper and lower cheek fullness scales, the same methodology was applied, using the additional subject photographs, again with 50 subjects being selected. Because only the upper and lower cheek fullness scales were of interest for this validation, all subjects of one validation cycle were assessed one after another in one booklet. The subject portraits included additional photographs to account for the newly introduced views (45° and chin down). The validation booklets with the updated cheek fullness scales were shipped to each expert for two validation cycles; the shipment of the second booklet took place 1 week after the rater had returned the first validation booklet.

For the reassessment, six of the 12 experts rated the 50 subjects using a scale version and pictures of the subjects displaying the face from two views (frontal and at 45°), whereas the remaining six experts rated the subjects using all three views (frontal, 45°, and chin down). The assessment of two views versus three views was assigned randomly.

Statistical Analyses

Descriptive statistics (arithmetic mean, standard deviation, standard error of the mean, minimum, 25% quartile, median, 75% quartile, maximum, and number of missing values) were calculated for the ratings per time point.

Reliability between raters (interrater reliability) and between the ratings of the first and second validation cycle (intrarater reliability) were evaluated to assess the reliability of the aesthetic scales. Interrater reliability on the assessments was evaluated using intraclass correlation coefficients (ICCs) at each time point of rating separately. The Shrout-Fleiss estimate for the ICC was used (assumption: the same raters, who are assumed to be a random subset of all possible raters, rate all subjects).[25,26] ICC of 0.20 or less is considered slight, 0.21 to 0.40 fair, 0.41 to 0.60 moderate, 0.61 to 0.80 substantial, and 0.81 and greater considered almost perfect.[27] The intrarater reliability of ratings was assessed using Pearson correlation coefficients for each scale and ICCs, because these are sensitive to a possible systematic bias in assessments. An intrarater reliability greater than 0.6 is considered high. The ICCs did not indicate any bias in the assessments, so only the Pearson correlation coefficients are discussed in this article. The inter- and intrarater reliability were assessed for all scales, the aesthetic areas, and the mid face aesthetic unit. The scores for each scale could range from 0 to 4. For the upper cheek aesthetic area (including infraorbital hollow and upper cheek fullness), the range of the summary score was 0 to 8; the range of the lower cheek aesthetic area summary score was 0 to 4. For the summary score of the mid face aesthetic unit the range was 0 to 12. No summary scores for the upper and lower cheek aesthetic areas were tested for the update of the cheek fullness scales, because these two scales were tested exclusively.

Results

There were 600 ratings in each of the two validation cycles (50 subjects × 12 experts) for each mid face scale.

Subject Characteristics

The mean age of the 50 subjects selected for the original validation (25 male, 25 female) was 51.7 ± 10.3 (range 25–66).

For validation of the updated upper and lower cheek fullness scales, subjects' mean age was 44.4 ± 12.1 (range 25–71). There were 35 women and 15 men.

Expert Characteristics

Of the 12 expert raters (9 male, 3 female), seven were dermatologists, three were researchers, one was an ophthalmologist, and one was a plastic surgeon.

TABLE 1. Descriptive Statistics for the Mid Face Rating Scales

Scale	Rating 1 Mean ± SD	Rating 1 Median (range)	Rating 2 Mean ± SD	Rating 2 Median (range)
Infraorbital hollow	1.57 ± 0.9	1.0 (0–4)	1.51 ± 0.9	1.0 (0–4)
Upper cheek fullness	1.61 ± 0.8	2.0 (0–4)	1.37 ± 0.8	1.0 (0–4)
Upper cheek fullness, updated				
Two views*	1.42 ± 1.0	1.0 (0–4)	1.35 ± 1.0	1.0 (0–4)
Three views†	1.69 ± 0.9	2.0 (0–4)	1.70 ± 1.1	2.0 (0–4)
Lower cheek fullness	0.91 ± 0.7	1.0 (0–3)	0.87 ± 0.6	1.0 (0–3)
Lower cheek fullness, updated				
Two views*	0.99 ± 1.0	1.0 (0–4)	0.99 ± 1.0	1.0 (0–4)
Three views†	1.60 ± 1.2	1.0 (0–4)	1.65 ± 1.2	2.0 (0–4)
Upper cheek aesthetic area summary score‡	3.18 ± 1.5	3.0 (0–8)	2.87 ± 1.5	3.0 (0–7)
Lower cheek aesthetic area summary score	0.91 ± 0.7	1.0 (0–3)	0.87 ± 0.6	1.0 (0–3)
Mid face aesthetic unit summary score	4.10 ± 1.8	4.0 (0–10)	3.75 ± 1.8	3.5 (0–8)

*Frontal and lateral (45°).
†Frontal, lateral (45°), and chin down.
‡Including infraorbital hollow and upper cheek fullness scores.
SD, standard deviation.

Validation of the Mid Face Scales

Descriptive statistics including the mean and median scores for the mid face scales, the two aesthetic areas and the mid face aesthetic unit are provided in Table 1. The mean scores for the infraorbital hollow scale were slightly higher at rating 1 compared to rating 2 (rating 1, 1.57; rating 2, 1.51), and the median scores were equal at both ratings (1.0).

For the updated upper cheek fullness scale, mean scores were higher for the three-view version (1, 1.69; 2, 1.70) than for the two-view version (1, 1.42; 2, 1.35) and were similar between the two rating cycles, in contrast to the original upper cheek fullness scale, which had a higher mean at cycle 1 (1, 1.61; 2, 1.37). Similarly, median scores were higher at rating cycle 1 (2.0) than rating cycle 2 (1.0) for the original upper cheek fullness scale, whereas for both versions of the updated scales, median scores were the same for the two cycles (two views: 1.0; three views: 2.0).

For the updated lower cheek fullness scale, mean scores were higher at both rating cycles (two views: 1 and 2, 0.99; three views: 1, 1.60, 2, 1.65) than for the original scale (1, 0.91; 2, 0.87). The median score was 1.0 for the original and updated lower cheek fullness scale (two views) at both rating cycles and 1.0 (1) and 2.0 (2) for the updated three-view scale. For the upper and lower cheek fullness scales, mean values were higher for the three-view version than for the two-view version.

The Shrout-Fleiss estimates for ICCs assessing the interrater reliability and Pearson correlation coefficients for test–retest reliability representing the intrarater reliability of validations are shown in Tables 2 and 3, respectively. The stability of ratings is visualized as bivariate scatter plots (bubble plots). Bubble plots of all experts pooled for each scale and the mid face summary score are shown in Figure 3.

Infraorbital Hollow Scale

Interrater reliability for the infraorbital hollow scale was substantial at both rating cycles (1, 0.72; 2, 0.66) (Table 2). Intrarater reliability of ratings was high (0.77, Table 3), also indicated by the

TABLE 2. Shrout-Fleiss Estimates for Intraclass Correlations Between Raters

Scale	Intraclass Correlation Coefficient	
	Rating 1	Rating 2
Infraorbital hollow	0.72	0.66
Upper cheek fullness	0.47	0.47
Upper cheek fullness, updated		
Two views*	0.67	0.66
Three views†	0.67	0.69
Lower cheek fullness	0.42	0.40
Lower cheek fullness, updated		
Two views*	0.67	0.74
Three views†	0.80	0.77
Aesthetic area		
Upper cheek‡	0.69	0.66
Lower cheek	0.42	0.40
Aesthetic unit: mid face	0.63	0.58

*Frontal and lateral (45°).
†Frontal, lateral (45°), and chin down.
‡Including infraorbital hollow and upper cheek fullness scores.

TABLE 3. Test–Retest Reliability

Scale	Correlation Coefficient (range)
Infraorbital hollow	0.77 (0.63–0.89)
Upper cheek fullness	0.65 (0.19–0.78)
Upper cheek fullness, updated	
Two views*	0.81 (0.70–0.85)
Three views†	0.83 (0.62–0.93)
Lower cheek fullness	0.57 (0.25–0.78)
Lower cheek fullness, updated	
Two views*	0.83 (0.80–0.92)
Three views†	0.88 (0.76–0.94)
Aesthetic area	
Upper cheek‡	0.77 (0.49–0.88)
Lower cheek	0.57 (0.25–0.78)
Aesthetic unit: mid face	0.71 (0.37–0.86)

*Frontal and lateral (45°).
†Frontal, lateral (45°), and chin down.
‡Including infraorbital hollow and upper cheek fullness scores.

bubbles located along the diagonal in the corresponding bubble plot (Figure 3A). The correlations were statistically significant ($P < .001$).

Upper Cheek Fullness Scale

Validation of the original upper cheek fullness scale yielded moderate interrater reliability at both cycles (both cycles: 0.47; Table 2). Intrarater reliability of ratings was high (0.65, Table 3). The range of intrarater reliability was large, indicating uncertainty in evaluating this scale. The correlations were statistically significant ($P < .001$).

Validation of the updated upper cheek fullness scale showed improvement (substantial interrater reliability for both scale versions; Table 2). Interrater reliabilities for three views and two views were the same at cycle 1 (0.67), whereas they were slightly higher for the three-view version than for the two-view version at cycle 2 (three views, 0.69; two views, 0.66). Intrarater reliability of ratings was high for both scale versions (Table 3). Intrarater reliability was slightly higher for the three-view version (0.83) than for the two-view version (0.81). The bubble plots show a more-diagonal distribution of the bubbles in both versions than for the original upper cheek fullness scale (Figures 3B–D). The correlations were statistically significant ($P < .001$).

Lower Cheek Fullness Scale

Similar to the upper cheek fullness scale, interrater reliability for validation of the original lower cheek fullness scale was moderate for the first cycle and fair for the second cycle (1, 0.42; 2, 0.40; Table 2).

Intrarater reliability of ratings was low for the original lower cheek fullness scale (0.57), as indicated by the location of bubbles in the corresponding bubble plot (Table 3, Figure 3E). The range of intrarater reliability was large, reflecting uncertainty in some of the experts' assessments of this scale. The correlations were statistically significant ($P < .001$).

Again, validation of the updated lower cheek fullness scale showed improvement (substantial interrater reliability for both scale versions; Table 2). Interrater reliability for the three-view version was higher than for the two-view version (three views: 1, 0.80; 2, 0.77; two views: 1, 0.67; 2, 0.74).

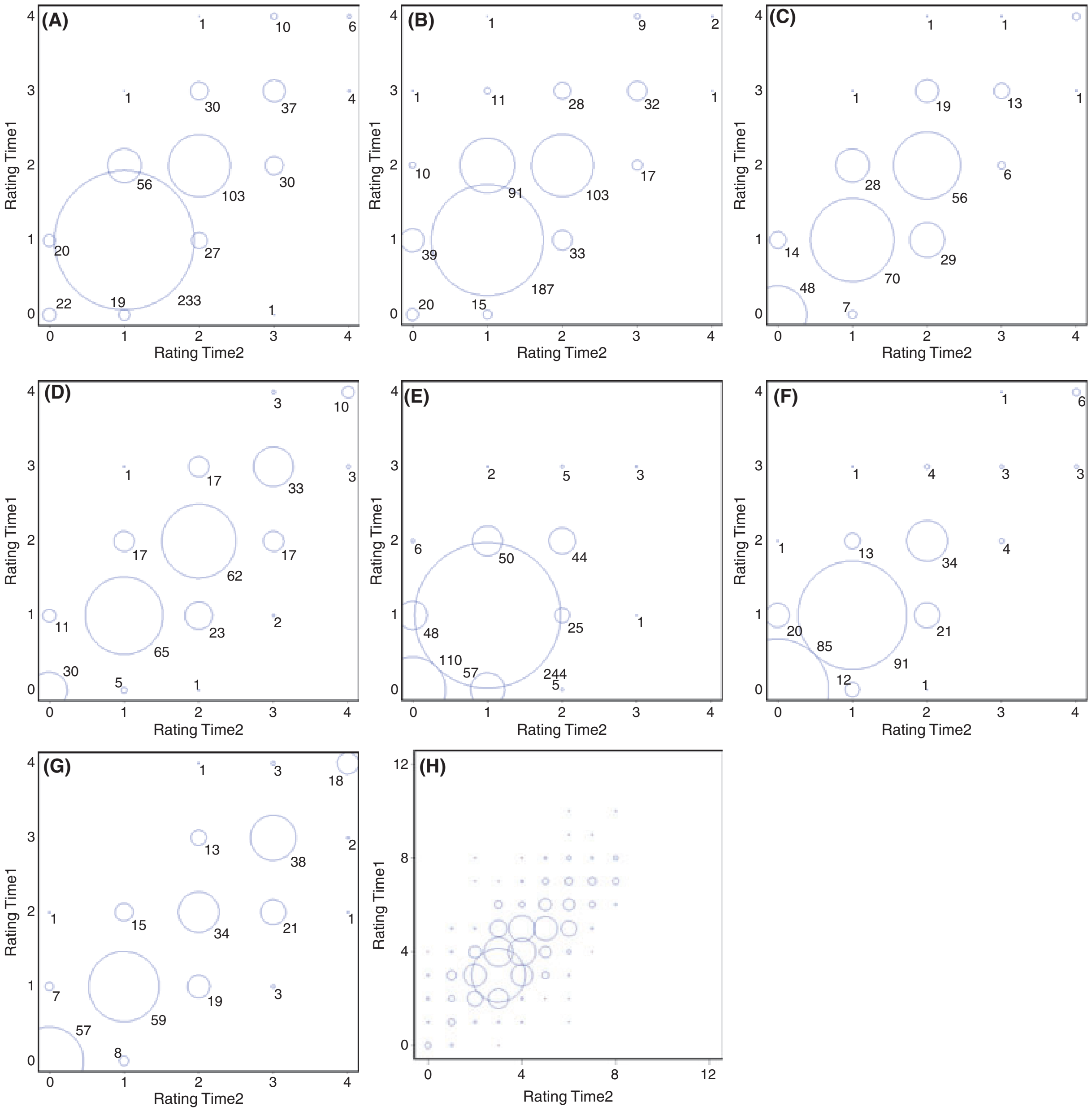

Figure 3. Bubble plots of rating combinations. A bubble plot represents the relationship between variables on a scatter plot; rating combinations between rating 1 (RatingTime1) and rating 2 (RatingTime2) are plotted using proportional circles to represent the frequencies of rating combinations. Hence, the plots illustrate intrarater reliability. Reliability is high if the bubbles are located along the diagonal and low if the bubbles are scattered randomly on the plot. The scale 0 to 4 on both axes represents the severity grades of the scales. The scale 0 to 12 on the axes of the bubble plot for the mid face aesthetic unit represents the summary score. (A) Infraorbital hollow. (B) Upper cheek fullness. (C) Upper cheek fullness, updated – two views. (D) Upper cheek fullness, updated – three views. (E) Lower cheek fullness. (F) Lower cheek fullness, updated – two views. (G) Lower cheek fullness, updated – three views. (H) Mid face aesthetic unit.

Intrarater reliability of ratings was high for both scale versions (Table 3). Similar to the upper cheek fullness scale, it was higher when three views were used (0.88) than with two views (0.83). The more-diagonal location of bubbles in the corresponding bubble plots also reflects the improvement in the updated scale (Figure 3F, G). The correlations were statistically significant (*P* < .001).

Summary Scores for the Upper and Lower Cheek Aesthetic Areas and Mid Face Aesthetic Unit

Interrater reliability for validation of the upper cheek aesthetic area summary score (including infraorbital hollow and the original upper cheek fullness scores) was substantial (1, 0.69; 2, 0.66; Table 2). For the lower cheek aesthetic area, which consists of the original lower cheek fullness scale only, interrater reliability was moderate for the first cycle (0.42) and fair for the second (0.40) (Table 2), as already mentioned. Interrater reliability for validation of the mid face aesthetic unit summary score was substantial for the first cycle (0.63) and moderate for the second (0.58). The bubble plot for the mid face aesthetic unit is shown in Figure 3H.

Intrarater reliability of ratings was high for the upper cheek aesthetic area summary score and mid face aesthetic unit (>0.70) and low for the lower cheek aesthetic area (0.57) (Table 3). All correlations were statistically significant (*P* < .001).

Discussion

The results of the mid face grading process within the framework of the Scale Summit II indicated varying interrater reliabilities, ranging from substantial for the infraorbital hollow scale to moderate for the original upper cheek fullness and fair to moderate for the original lower cheek fullness, depending on the rating cycle. Intrarater reliability was shown to be high for the infraorbital hollow and the original upper cheek fullness scale but low for the original lower cheek fullness scale.

The development of the original upper and lower cheek fullness scales was based on considerable effort but resulted in only fair to moderate interrater reliabilities. With such results, no recommendation on the future use of the cheek fullness scales can be provided. The rating of fullness changes may have been difficult for the raters because the three-dimensional aspect of fullness may not have been captured sufficiently in the two-dimensional photographs, particularly with only one frontal view. Furthermore, splitting the mid face into individual mid face scales that did not respect the fat compartmentalization (medial, middle, and lateral temporal cheek fat) may be a possible explanation of the different interrater reliabilities obtained for the original cheek fullness scales. In the upper cheek fullness scale, changes of the more central part of the upper cheek are presented, but the high variance of the zygoma may have complicated the rating assessment. Regarding the lower cheek fullness scale, the more-severe gradings were not fully represented in the validation population because Grade 4 of the lower cheek fullness scale (very severely sunken lower cheek) is not common in the normal population. Therefore, a clear floor effect was present that might have affected reliability.

Because the reliability results for the original cheek fullness scales could not be considered satisfactory for this important area and would have excluded the use of the scales in the routine clinical setting, a more specialized approach was adopted for improvement of the upper and lower cheek fullness scales. The combination of up to three different views in one scale was subsequently used to address the three-dimensional aspect of volume deficiencies. This methodology led to considerably higher reliabilities when the two scales were re-assessed; interrater reliability was substantial. Because more attention was paid to the three-dimensionality of the areas assessed and to a more even distribution of the severity grades among the population, a more uniform assessment could be achieved. Intrarater reliability was also high for the updated upper and lower cheek fullness scales. The two ratings took place within 4 weeks of each other—an interval long enough to reduce memory effects and likely to reflect common practice for re-evaluating or following up of patients in the aesthetic clinic.

Overall, higher inter- and intrarater reliabilities were obtained when the mid face was viewed with the additional chin-down perspective than with the two-view version with frontal and 45º views only. Furthermore, the results (in particular the bubble plots) illustrate that using all three views results in broader, more specific use of the severity grades when rating the subjects, than the two-view version. Thus, the three-view approach allows a more sensitive assessment of the three-dimensionality of the cheek and might better reflect the real-life situation in which a physician examines an individual patient from more than two angles.

Optimal mid face rating scales are important and increasingly indispensable tools in physician–patient communication regarding short- and long-term outcomes of aesthetic mid face treatments. By using the aesthetic scales, the focus is shifted from specific and sometimes long medical explanations, conceptional drawings, or single case photographs to photographs of one subject in which only the trait of interest is changed. Our experience from the Scale Summit I[16–21] has already shown that such scales can facilitate discussions between physicians and patients on age-related changes in the face.

These scales were validated and are limited to a population aged 25 to 66 with Fitzpatrick skin types I through IV and mainly with no previous facial aesthetic treatment. Consequently, next steps for these scales should aim at their application in daily practice to further investigate their sensitivity and construct validity. Assessing aesthetic treatment changes over time or testing populations with different qualitative traits (e.g., different ethnicities, facial types, or mechanisms of aging) could be possible research scenarios for such questions. More specifically, it may be of interest to use the ever more available and popular three-dimensional technology for presentation of subjects in these and other rating scales. Using booklets with the usual letter format print-outs of the aesthetic units for the ratings of the Scale Summit II was a methodology decision towards an easier application in daily practice and repeatability. Highly magnified and larger presentations of the single traits or three-dimensional technology might have resulted in a methodology too specific for routine use.

In conclusion, the mid face scales with the updated cheek fullness scales described here represent a major contribution in aesthetic medicine as valid and reliable tools for the assessment of age-related changes.

References

1. Shaw RB Jr, Kahn DM. Aging of the midface bony elements: a three-dimensional computed tomographic study. Plast Reconstr Surg 2007;119:675–81.
2. Carruthers J, Glogau RG, Blitzer A, Facial Aesthetics Consensus Group Faculty. Advances in facial rejuvenation: botulinum toxin type A, hyaluronic acid dermal fillers, and combination therapies - consensus recommendations. Plast Reconstr Surg 2008;121:S5–30.
3. Mendelson BC, Hartley W, Scott M, McNab A, Granzow JW Age-related changes to the orbit and midcheek and the implications for facial rejuvenation. Aesthetic Plast Surg 2007;31:419–23.
4. Tan SR, Glogau RG. Filler esthetics. In: Carruthers A, Carruthers J, editors. Procedures in cosmetic dermatology series: Soft tissue augmentation. Philadelphia: Saunders; 2005. pp. 11–8.
5. Brandt FS, Cazzaniga A. Hyaluronic acid gel fillers in the management of facial aging. Clin Interv Aging 2008;3: 153–9.
6. Rohrich RJ, Pessa JE. The fat compartments of the face: anatomy and clinical implications for cosmetic surgery. Plast Reconstr Surg 2007;119:2219–27; discussion 2228-31.
7. Rohrich RJ, Arbique GM, Wong C, Brown S, et al. The anatomy of suborbicularis fat: implications for periorbital rejuvenation. Plast Reconstr Surg 2009;124:946–51.
8. Schaverien MV, Pessa JE, Rohrich RJ. Vascularized membranes determine the anatomical boundaries of the subcutaneous fat compartments. Plast Reconstr Surg 2009;123:695–700.
9. James J, Carruthers A, Carruthers J. HIV-associated facial lipoatrophy. Dermatol Surg 2002;28:979–86.
10. Coleman S, Saboeiro A, Sengelmann R. A comparison of lipoatrophy and aging: volume deficits in the face. Aesthetic Plast Surg 2009;33:14–21.
11. Bowler PJ. Impact of facial rejuvenation with dermatological preparations. Clin Interv Aging 2009;4:81–9.
12. Carruthers J, Carruthers A. Facial sculpting and tissue augmentation. Dermatol Surg 2005;31:1604–12.

13. Hirsch RJ, Carruthers J, Carruthers A. Infraorbital hollow treatment by dermal fillers. Dermatol Surg 2007;33:1116–9.

14. Rhee JS, McMullin BT. Outcome measures in facial plastic surgery. Patient-reported and clinical efficacy measures. Arch Facial Plast Surg 2008;10:194–207.

15. Rzany B, Carruthers A, Carruthers J, Flynn TC, et al. Validated composite assessment scales for the global face. Dermatol Surg 2012;38:295–309.

16. Carruthers A, Carruters J, Hardas B, Kaur M, et al. A validated brow positioning grading scale. Dermatol Surg 2008;34:S150–4.

17. Carruthers A, Carruters J, Hardas B, Kaur M, et al. A validated grading scale for forehead lines. Dermatol Surg 2008;34:S155–60.

18. Carruthers A, Carruthers J, Hardas B, Kaur M, et al. A validated lip fullness grading scale. Dermatol Surg 2008;34:S161–6.

19. Carruthers A, Carruthers J, Hardas B, Kaur M, et al. A validated grading scale for marionette lines. Dermatol Surg 2008;34:S167–72.

20. Carruthers A, Carruters J, Hardas B, Kaur M, et al. A validated grading scale for crow's feet. Dermatol Surg 2008;34:S173–8.

21. Carruthers A, Carruters J, Hardas B, Kaur M, et al. A validated hand grading scale. Dermatol Surg 2008;34:S179–83.

22. Shoshani D, Markovitz E, Monstrey SJ, Narins DJ. The modified Fitzpatrick Wrinkle Scale: a clinical validated measurement tool for nasolabial wrinkle severity assessment. Dermatol Surg 2008;34:S85–9.

23. Monheit GD, Gendler EC, Poff B, et al. Development and validation of a 6-point grading scale in patients undergoing correction of nasolabial folds with a collagen implant. Dermatol Surg 2010;36(Suppl 3):1809–16.

24. Hund T, Ascher B, Rzany B for the SMILE Study Group. Reproducibility of two-four-point clinical scores for lateral canthal lines (crow's feet). Dermatol Surg 2006;32: 1256–60.

25. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420–8.

26. Yaffee RA. Enhancement of reliability analyses: application of intraclass correlations with SPSS/Windows v.8. New York: Statistics and Social Science Group; 1998. http://www.nyu.edu/its/socsci/Docs/intracls.html.

27. Landis J, Koch G. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.

Address correspondence and reprint requests to:
Jean Carruthers, MD, 943 West Broadway, Suite 820,
Vancouver, British Columbia, V5Z4E1, Canada, or
e-mail: drjean@carruthers.net