OPEN

Validation of Photonumeric Assessment Scales for Temple Volume Deficit, Infraorbital Hollows, and Chin Retrusion

Amir Moradi, MD,* Xiaoming Lin, RN, MS,† Shawn Allen, MD,‡ Steven Fagien, MD,§ Maria Norberg, PhD,† and Stacy Smith, MD‖

BACKGROUND Assessment scales are valuable tools in aesthetic clinical research and practice.

OBJECTIVE To validate 3 photonumeric scales covering temple volume deficit, infraorbital hollows, and chin retrusion.

MATERIALS AND METHODS Subjects reflecting the whole range of the scales were assessed independently by 3 evaluators at 2 separate occasions. Intraobserver agreement (the ability of each evaluator to assess the same grade for a specific subject at both evaluation occasions) and interobserver agreement (the degree to which evaluators independently provided identical grades for the same subject) were measured by weighted kappa statistics and percent exact agreement.

RESULTS Approximately 70 subjects were included in each scale validation. The predefined success criteria of an intraobserver weighted kappa coefficient of ≥0.6 and an interobserver median pairwise weighted kappa coefficient of ≥0.6 were met for each scale. These results indicate substantial agreement, both between the 2 evaluations, and between the 3 evaluators.

CONCLUSION These scales covering temple volume deficit, infraorbital hollows, and chin retrusion are validated assessment tools, based on live evaluations. Intraobserver agreement (between the 2 evaluations) and interobserver agreement (between the 3 evaluators) were both substantial.

The study was funded by Galderma. A. Moradi is a consultant, and clinical investigator for Galderma. Lin is employed by Galderma. S. Allen is a consultant trainer and advisor for Galderma. S. Fagien is a consultant, advisor, and clinical investigator for Galderma. M. Norberg is employed by Galderma. S. Smith is a consultant and clinical investigator for Galderma.

Validated assessment scales are versatile tools, used by aesthetic health care practitioners to communicate treatment goals with patients, showing where the patient currently is regarding an indication, and what the results could be after augmentation. Scales are also valuable for standardized assessments of treatment effect in clinical studies. There are assessment scales available for various indications including but not limited to upper facial lines,[1–3] temple volume deficit,[4] infraorbital hollows,[5] nasolabial folds,[6,7] midface,[8–10] marionette lines,[11] lip fulness,[12] hand wrinkles,[13,14] and chin.[15] The objective of this study was to validate 3 new 4-grade photonumeric assessment scales using live evaluations; the Galderma Temple Volume Deficit Scale (GTVDS), Galderma Infraorbital Hollows Scale (GIHS), and Galderma Chin Retrusion Scale (GCRS)

Private Practice, Vista, California; †Galderma Aesthetics, Uppsala, Sweden; ‡Dermatology Specialists of Boulder, Boulder, Colorado; §Private Practice, Boca Raton, Florida; ‖Private Practice, Encinitas, California

ISSN: 1076-0512 • Dermatol Surg 2020;46:1148–1154 • DOI: 10.1097/DSS.0000000000002269

(Figures 1–3). The grades of the scales represent visibly distinct degrees of deficiency/retrusion where 0 = none; 1 = mild; 2 = moderate; 3 = severe.

Material and Methods

Study Design

The scales were developed first by selecting a subject position and lighting to best demonstrate the anatomic difference under assessment. Subsequently, representative example photographs were selected from a larger pool of images captured by Canfield Scientific, INC. (Fairfield, NJ). Several examples of each grade were included along with one set of images (morphed) that were created by editing one image into greater and lesser degrees of severity for that finding.

This was a live scale validation study to evaluate intraobserver and interobserver agreement of the scales. Intraobserver agreement referred to the ability of each evaluator to reproduce their score for a specific subject from the first evaluation at the second evaluation, and interobserver agreement was the degree to which evaluators independently provided identical grades for the same subject. Enrolled subjects were assessed independently by 3 evaluators on 2 separate occasions, 13 to 14 days apart to reduce the risk of memory bias.

At the first assessment, subjects were presented to the evaluators according to a unique subject number that was assigned to them in consecutive order as they arrived to the validation event. At the second assessment, the subjects were evaluated in a randomized order.

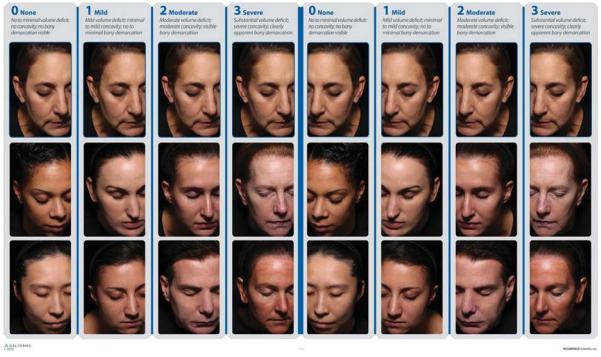

Figure 1. Galderma Temple Volume Deficit Scale. © 2019, Galderma. All rights reserved.

To ensure data quality and consistency of the evaluations, all evaluators received training on the use of the scales before the first assessment. Assessments were made individually and were at no point discussed between the evaluators. Extreme care was taken to standardize the lightning and positioning of the subjects. The results were recorded in electronic case report software. The investigators assisting in development of the scales, that is, the authors of this publication, were not part of the validation assessments. Institutional review board approval was not required as study procedures were limited to obtaining photographic images and performing live evaluations (i.e., minimal risk). All subjects signed a photographic release form.

Study Assessments

In addition to the live scale assessments, demographic data were collected, and subject photographs were taken for documentation purposes.

Statistical Methods

The sample size of these validations was not based on any statistical considerations; the aim was to include heterogeneous subject groups representing different races, skin types, sex, and age, and approximately equal proportions of each grade of severity within the scales.

All statistical analyses were performed using the SAS version 9.4 statistical software package. Consistency in quantitative scale assessments was determined by intraobserver and interobserver agreement evaluated using weighted kappa coefficient statistics according to Cicchetti and Allison (with their associated 2-sided 95% confidence intervals [CIs]) as well as percentage of exact agreement. For the scales to be valid, an overall intraobserver weighted kappa coefficient of ≥ 0.6 and an interobserver median pairwise weighted kappa coefficient of ≥ 0.6 were set as success criteria. The analyses of intraobserver and interobserver

Figure 2. Galderma Infraorbital Hollows Scale. © 2019, Galderma. All rights reserved.

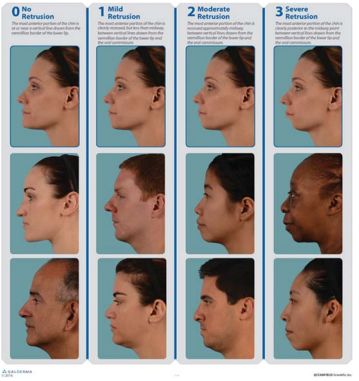

Figure 3. Galderma Chin Retrusion Scale. © 2019, Galderma. All rights reserved.

agreement were repeated within subgroups determined by race. For ease of interpretation and discussion of the results, labels for different ranges of kappa coefficient statistics were used, as shown in Table 1. Demographic characteristics were analyzed descriptively.

Intraobserver Agreement

The intraobserver agreement was evaluated by comparison of the initial and the replicate evaluations of the subjects for each evaluator. Overall percentage of exact agreement was calculated for the complete set of data by dividing the sum of the number of observations (GIHS: both right and left sides of the face) with equal ratings at both occasions with the total number of observations. Percentage of exact agreement was also calculated separately for each evaluator. Weighted kappa coefficients were calculated for each evaluator separately (GIHS: calculated across sides) and were also stratified by evaluator to obtain estimates for the overall weighted kappa coefficient (GIHS: calculated across evaluators and sides).

Interobserver Agreement

Interobserver agreement was evaluated by comparing each of the pairs of evaluators from both occasions. Percentage of exact agreement for each pair of evaluators was calculated by dividing the sum of all observations (GIHS: both right and left sides of the face) with equal ratings with the total number of observations. This was calculated across occasions, as well as separately for each occasion. Similarly, weighted kappa coefficients were calculated for each pair of evaluators separately for each occasion (weighted kappa across sides of the face) and were also stratified by occasion to obtain estimates for the overall weighted kappa coefficients for each pair of evaluators (GIHS: weighted kappa across occasions and sides of the face).

Results

The evaluations for GTVDS and GCRS took place 13 days apart, and evaluations for GIHS were held 14 days apart.

Subject Demographics

Demographic characteristics of the subjects who were included in these scale validations are summarized in Table 2. The mean age of the subjects was 42 to 48 years, with an overall age range of 20 to 85 years. The majority of the subjects were women (57%–71%) and white (60%–88%). Because there were only a few subjects representing certain races, some race groups were pooled in the subgroup analyses determined by race (Table 2); for GTVDS, there were 63 (87.5%) white subjects, and 9 (12.5%) subjects in the Pooled group; for GIHS, there were 54 (76%) white subjects, and 17 (24%) subjects in the Pooled group. For GCRS, there were 43 (60%) white subjects, 14 (19%) Asian subjects, and 15 (21%) subjects in the Pooled group.

Intraobserver Agreement

Overall intraobserver agreement between the 2 evaluations was substantial for all 3 scales. The

TABLE 1. Agreement of Weighted Kappa Coefficients

Kappa Coefficient	*Strength of Agreement*
0.00–0.19	Poor agreement
0.20–0.39	Fair agreement
0.40–0.59	Moderate agreement
0.60–0.79	Substantial agreement
0.80–1.00	Almost perfect agreement

predefined success criterion of an overall intraobserver weighted kappa coefficient of ≥0.6 was thus met for each scale. In addition, individual agreement between results from the 2 evaluations was also substantial for each evaluator, also covering all scales.

Temple Volume Deficit Scale
For GTVDS, overall weighted kappa coefficient was 0.74 (95% CI: 0.69–0.78), and overall exact agreement was 70%. Individual weighted kappa coefficients for the 3 evaluators ranged from 0.69 to 0.77; individual exact agreement ranged from 66% to 72% (Figure 4).

Infraorbital Hollows Scale
For GIHS, the overall weighted kappa coefficient was 0.73 (95% CI: 0.68–0.78). Overall exact agreement was 72%. Individual weighted kappa coefficients for the 3 evaluators ranged from 0.69 to 0.75, and individual exact agreement ranged from 72% to 73% (Figure 4).

Chin Retrusion Scale
Overall weighted kappa coefficient for intraobserver agreement was 0.70 (95% CI: 0.64–0.77). Overall exact agreement was 71%. Individual weighted kappa coefficients for the 3 evaluators ranged from 0.65 to 0.73, individual exact agreement ranged from 63% to 76% (Figure 4).

Intraobserver Subgroup Analysis by Race

Intraobserver agreement between the 2 evaluations was substantial for all scales also when analyzed by race groups. The predefined success criterion of an overall intraobserver weighted kappa coefficient of ≥0.6 was met for all race categories, showing no evident difference in intraobserver reliability.

Temple Volume Deficit Scale
The overall weighted kappa coefficients for the white and Pooled groups were 0.72 and 0.85, respectively, that is, well above 0.6 for both categories. The exact agreement was 68% for white subjects and 80% for Pooled subjects (Figure 5).

Infraorbital Hollows Scale
For GIHS, the overall weighted kappa coefficients for white and Pooled subjects were 0.67 and 0.69, respectively. Exact agreement was 74% for white subjects and 68% for Pooled subjects (Figure 5).

Chin Retrusion Scale
Overall weighted kappa coefficients were similar between the 3 groups: white, Asian, and the Pooled group, ranging from 0.66 to 0.71. Exact agreement ranged from 67% to 74% (Figure 5).

Interobserver Agreement

Interobserver kappa coefficients and exact agreement were calculated for each pair of evaluators from both occasions. Evaluator pairs were not the same for all 3 scale validations. The results indicated substantial agreement between the 3 pairs of evaluators for all scales.

Temple Volume Deficit Scale
The median kappa coefficients of the 3 pairs of evaluators were 0.72, well above the predefined acceptable value of 0.6. The exact agreement for the 3 pairs of evaluators ranged from 68% to 69% (Figure 6).

Infraorbital Hollows Scale
The median kappa coefficient was 0.67. The exact agreement for the 3 pairs of evaluators ranged from 63% to 69% (Figure 6).

TABLE 2. Demographic Characteristics

Variable	*Parameter*	*GTVDS* n = *72**	*GIHS* n = *71†*	*GCRS* n = *72‡*
Age, yr	Mean	48.3	42.6	41.9
	SD	13.4	14.3	14.7
	Median	50.0	41.0	37.0
	Minimum	24	22	20
	Maximum	85	73	70
Sex, *n* (%)	Women	51 (71)	42 (59)	41 (57)
	Men	21 (29)	29 (41)	31 (43)
Fitzpatrick skin type, *n* (%)	I	14 (19)	2 (3)	5 (7)
	II	25 (35)	20 (28)	15 (21)
	III	23 (32)	24 (34)	25 (35)
	IV	5 (7)	23 (32)	8 (11)
	V	3 (4)	2 (3)	14 (19)
	VI	2 (3)	0 (0)	5 (7)
Race, *n* (%)	Other	1§ (1)	4§ (6)	—
	Black or African American	1§ (1)	2§ (3)	5§ (7)
	Hispanic or Latin	5§ (7)	—	10§ (14)
	White	63 (88)	54 (76)	43 (60)
	Asian	2§ (3)	11§ (15)	14 (19)

*Seventy-two and 69 subjects were assessed at first and second evaluation occasions, respectively.
†Seventy-one and 67 subjects were assessed at first and second evaluation occasions, respectively.
‡All 72 subjects were assessed at both evaluation occasions.
§Pooled subjects in subgroup analyses.
GTVDS, Galderma Temple Volume Deficit Scale; GIHS, Galderma Infraorbital Hollows Scale; GCRS, Galderma Chin Retrusion Scale.

Chin Retrusion Scale

For GCRS, the median weighted kappa coefficient was 0.70. Exact agreement between the 3 pairs of evaluators ranged from 69% to 76% (Figure 6).

Interobserver Subgroup Analysis by Race

Temple Volume Deficit Scale

The median weighted kappa coefficients was 0.72 for the white group, and 0.68 for the Pooled group; the predefined success criterion of a median pairwise interobserver weighted kappa coefficient of ≥0.6 was thus met. Exact agreement between the 3 pairs of evaluators was ≥67% in the white group and ≥64% in the Pooled group (not shown).

Infraorbital Hollows Scale

The median weighted kappa coefficients were 0.65 for the white group, and 0.55 for the Pooled group, that is, slightly higher for the white group. Exact agreement between the 3 pairs of evaluators was ≥67% in the white group and ≥49% in the Pooled group (not shown).

Chin Retrusion Scale

The median weighted kappa coefficients were 0.72 for the white group, 0.70 for the Asian group, and 0.63 for the Pooled group. The predefined success

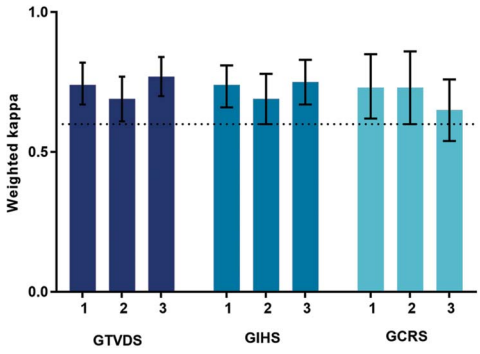

Figure 4. Intraobserver agreement by evaluator. Intraobserver agreement by evaluator displayed as weighted kappa with 95% confidence interval. Dashed line represents the predefined success criteria of an intraobserver weighted kappa coefficient of ≥0.6.

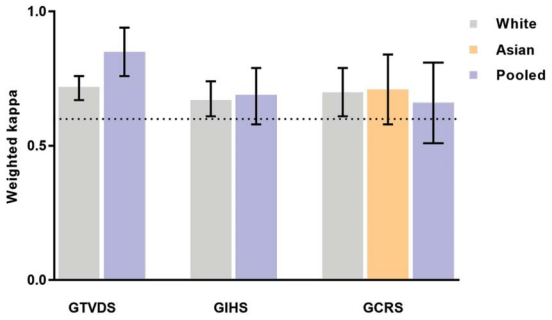

Figure 5. Intraobserver agreement by race group. Intraobserver agreement by race group displayed as weighted kappa. Dashed line represents the predefined success criteria of an intraobserver weighted kappa coefficient of ≥0.6.

criterion of a median pairwise interobserver weighted kappa coefficient of ≥0.6 was thus met also for each race group. Exact agreement between the 3 pairs of evaluators ranged from 74% to 76% in the white group, from 64% to 82% in the Asian group, and from 57% to 73% in the Pooled group (not shown).

Discussion

This was a live scale validation study to evaluate intraobserver and interobserver agreement of 3 assessment scales covering temple volume deficit, infraorbital hollows, and chin retrusion. The 4 grades of the scales represent clinically distinct steps, and the facial assessment areas are defined by photographs and accompanying text descriptions for a clear interpretation of each grade. To allow an overall assessment of the face, the scale photographs display the full face of the subjects, and not only the area of concern. According to the scale developers, the grading process was facilitated by a vertical presentation of facial images displaying the same grade on the scales, hence providing the possibility to move the patient horizontally between grades on the scales. In addition, the morphed photographs provide the evaluator with the possibility to focus exclusively on differences in appearance between each grade on the scale and to avoid distractions of other differences in each grade example. The 2 additional unaltered photographs per scale step show the attributes for each grade in individuals with different sex and ethnicity. For GTVDS and GIHS, both sides of the face are displayed, providing ability to assess both sides of the face; mirrored images of the 2 sides of the face were chosen over actual right and left sides from the subjects for consistency. An additional notable aspect of the GCRS is a unique assessment area, using the oral commissure and lower vermilion border as consistent landmarks.

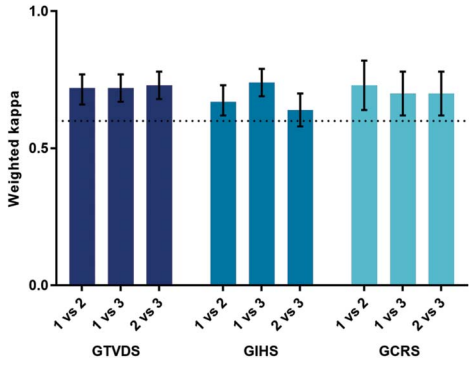

Figure 6. Interobserver agreement. Interobserver agreement displayed as weighted kappa. Dashed line represents the predefined success criteria of the median of the pairwise weighted kappa coefficients ≥0.6.

All validations were based on live assessments, which are preferable over photographic assessments, as this resource investment provides scientific merit and makes the scales suitable for use on live subjects.

Intraobserver agreement, that is, the ability of each evaluator to reproduce their original score at a second assessment occasion, was evaluated using weighted kappa coefficients and associated categorical grading for interpretation guidance. The intraobserver weighted kappa coefficient values stratified by evaluator were above the predefined success criterion of an intraobserver weighted kappa coefficient of at least 0.6 and show that for all scales, and for each evaluator, there was substantial agreement between the first and second evaluation ratings for the same subject. The overall exact agreement (63%–76%) was consistent with this result.

Interobserver agreement, that is, the ability of the evaluators to independent of each other assign the same subject an identical score, was evaluated by weighted kappa coefficients and exact agreement for each pair of evaluators. The median weighted

kappa coefficients for interobserver agreement were above 0.6, which shows that there was substantial agreement between the evaluators.

Intraobserver and interobserver agreement was substantial also when analyzed by race category and met the predefined success criteria for all scales, except for GIHS where interobserver reliability was slightly higher in the larger white group, which can be explained by the low number of individuals in the Pooled group (n = 17 [24%]).

Limitations with these validation studies include that the number of subjects was not evenly distributed between different race categories. Furthermore, an even more heterogeneous subject population would have been preferred to strengthen the generalizability of the scales. To ensure data quality and consistency in evaluations when using the scales, it is also essential that health care professionals receive adequate training on the use of the scales before implementation in clinical routine and clinical studies.

Conclusions

Intraobserver and interobserver agreement for these 3 scales covering temple volume deficit, infraorbital hollows, and chin retrusion was substantial, providing reliable and reproducible results both when used by different evaluators and by the same evaluator on different occasions. As the predefined success criteria were met, the scales are thus validated assessment tools, suitable for use in clinical practice to facilitate communication of treatment goals with patients, and in clinical studies evaluating treatment effect.

Acknowledgments The authors thank Joely Kaufman-Janette, MD, Lisa Donofrio, MD, Charlie Finn, MD, and Z. Paul Lorenc, MD, for serving as evaluators in this scale validation study. Patients provided written consent for the use of their images.

References

1. Honeck P, Weiss C, Sterry W, Rzany B. Reproducibility of a four-point clinical severity score for glabellar frown lines. Br J Dermatol 2003;149:306–10.
2. Carruthers A, Carruthers J, Hardas B, Kaur M, et al. A validated grading scale for crow's feet. Dermatol Surg 2008;34(Suppl 2):S173–178.
3. Carruthers A, Carruthers J, Hardas B, Kaur M, et al. A validated grading scale for forehead lines. Dermatol Surg 2008;34(Suppl 2):S155–160.
4. Carruthers J, Jones D, Hardas B, Murphy DK, et al. Development and validation of a photonumeric scale for evaluation of volume deficit of the temple. Dermatol Surg 2016;42(Suppl 1):S203–s210.
5. Donofrio L, Carruthers J, Hardas B, Murphy DK, et al. Development and validation of a photonumeric scale for evaluation of infraorbital hollows. Dermatol Surg 2016;42(Suppl 1):S251–s258.
6. Buchner L, Vamvakias G, Rom D. Validation of a photonumeric wrinkle assessment scale for assessing nasolabial fold wrinkles. Plast Reconstr Surg 2010;126:596–601.
7. Day DJ, Littler CM, Swift RW, Gottlieb S. The wrinkle severity rating scale: a validation study. Am J Clin Dermatol 2004;5:49–52.
8. Lorenc ZP, Bank D, Kane M, Lin X, et al. Validation of a four-point photographic scale for the assessment of midface volume loss and/or contour deficiency. Plast Reconstr Surg 2012;130:1330–6.
9. Carruthers J, Flynn TC, Geister TL, Görtelmeyer R, et al. Validated assessment scales for the mid face. Dermatol Surg 2012;38:320–32.
10. Narins RS, Carruthers J, Flynn TC, Geister TL, et al. Validated assessment scales for the lower face. Dermatol Surg. 2012;38:333–42.
11. Carruthers A, Carruthers J, Hardas B, Kaur M, et al. A validated grading scale for marionette lines. Dermatol Surg 2008;34(Suppl 2):S167–172.
12. Carruthers A, Carruthers J, Hardas B, Kaur M, et al. A validated lip fullness grading scale. Dermatol Surg 2008;34(Suppl 2):S161–166.
13. Carruthers A, Carruthers J, Hardas B, Kaur M, et al. A validated hand grading scale. Dermatol Surg 2008;34(Suppl 2):S179–183.
14. Cohen JL, Carruthers A, Jones DH, Narurkar VA, et al. A randomized, blinded study to validate the merz hand grading scale for use in live assessments. Dermatol Surg 2015;41(Suppl 1):S384–388.
15. Sykes JM, Carruthers A, Hardas B, Murphy DK, et al. Development and validation of a photonumeric scale for assessment of chin retrusion. Dermatol Surg 2016;42(Suppl 1):S211–s218.

Address correspondence and reprint requests to: Amir Moradi, MD, Private Practice, 2023 W. Vista Way, Suite F, Vista, CA 92083, or e-mail: moradimd@gmail.com