The Modified Fitzpatrick Wrinkle Scale: A Clinical Validated Measurement Tool for Nasolabial Wrinkle Severity Assessment

David Shoshani, MD,* Elana Markovitz, RN,* Stan J. Monstrey, MD, PhD,† and David J. Narins, MD, FACS‡

BACKGROUND The number of existing wrinkle assessment scales makes it difficult to compare the efficacy of cosmetic techniques in rejuvenating photoaged skin. A single and simple assessment scale that reliably quantifies wrinkle depth is needed.

OBJECTIVE The objective was to validate the Modified Fitzpatrick Wrinkle Scale (MFWS) as a nasolabial wrinkle severity assessment tool.

METHODS AND MATERIALS The MFWS comprises three main classes, in which definitions are based on a set of reference photographs and descriptions, and three interclasses, in which definitions are based only on descriptions. Assessors were trained to apply this scale to volunteers and study patients by using photographs of nasolabial wrinkles either alone or with descriptions. Inter- and intraassessment reliability coefficients were calculated using weighted kappa statistics.

RESULTS In patients, the combined intraassessor reliability from both sides of the face was 0.71 (95% confidence interval [CI], 0.68–0.74) when only photographs were used and 0.79 (95% CI, 0.76–0.82) when descriptions were added. Interassessor reliability for the photographs alone was 0.65 (95% CI, 0.62–0.68) and 0.74 (95% CI, 0.69–0.79) for photographs plus descriptions.

CONCLUSIONS The MFWS is a reliable method for quantitative assessment of nasolabial skin folds, with good inter- and intraassessor reliability. Including descriptions with the photographs increased reliability.

This study was funded by Colbar Lifescience Ltd, Israel.

Demand for rejuvenation of photoaged skin is increasing, and thus the need to assess treatment outcomes has become more important. Over the years, a variety of assessment systems to measure the severity of wrinkles have been proposed. Many of these systems have proved useful in assessing diverse skin aging processes such as smoking-associated facial wrinkling in young people[1] and photoaging,[2,3] as well as assessing various treatment options such as wrinkle-improving lipstick.[4] However, they depend on the availability of sophisticated imaging equipment and technology such as photoimaging, high-frequency ultrasonography, and more recently, multiphoton fluorescence and second-harmonic-generation microscopy.

Simpler wrinkle severity evaluation systems have been produced that rely on comparisons of photographs. Although more subjective, these methods are popular among clinicians. Wrinkle grading systems of varying complexity have been validated for reproducibility and reliability and are used for assessing the efficacy of treatments such as botulinum toxin A injections[5,6] and hormone replacement therapy[7] and for the classification of facial wrinkles.[8] Other examples include the Wrinkle Severity Rating Scale (WSRS), which is a 5-grade assessment system of labial folds that was validated[9] and then applied in two studies to distinguish between two treatments for facial soft tissue augmentation,[10,11] and the Lemperle scale, which was used in a study

*ColBar LifeScience, Herzliya, Israel; †Department of Plastic Surgery, University Hospital Gent, Gent, Belgium; ‡New York University School of Medicine, New York, New York

ISSN: 1076-0512 • Dermatol Surg 2008;34:S85–S91 • DOI: 10.1111/j.1524-4725.2008.34248.x

(abstract presented in AAD, 2006,[12] to compare two treatments for nasolabial fold correction.

With the growing number of wrinkle rating systems, evaluating the efficacy of different treatments between studies is becoming increasingly difficult. Thus, there is a need for a single, standardized, objective, and reliable method for measuring the severity of facial wrinkles and folds to evaluate and compare the efficacy of cosmetic treatments. In 1996, Fitzpatrick and coworkers[13] proposed a wrinkle-scoring system for assessing perioral and periorbital wrinkle severity in a study evaluating the efficacy of laser treatment in resurfacing photoaged skin. This classification was based on generalized wrinkling, elastosis, and dyschromia as well as wrinkle depth. Using reference photographs, the wrinkles were classified into one of the three classes (1, 2, or 3), which were defined as mild, moderate, or severe. Instead of interclasses, each of the three defined classes provided an additional three subscores; however, these subscores were represented by a typical photograph. This system for defining skin type wrinkles was subsequently used in numerous trials to demonstrate improvements in patients receiving treatment for photoaged skin.[14–16]

In this study, we used a Modified Fitzpatrick Wrinkle Scale (MFWS) for the assessment of nasolabial folds. The four main classes of wrinkle severity were defined based on photography and descriptors. Instead of subscores, the MFWS included three additional interclasses, which were defined based on descriptions alone. The objective of the study was to determine the reproducibility and reliability of the MFWS as a clinical measurement tool for assessment of nasolabial wrinkle severity in volunteer and clinical study populations.

Methods and Materials

This validation study was carried out at the Medical Department of ColBar LifeScience and used photographs from volunteers and clinical study patients undergoing treatment for nasolabial wrinkles.[17] The study was conducted in accordance with the International Conference on Harmonisation Guidelines for Good Clinical Practice. All patients signed the informed consent form.

Classes of MWFS

The MFWS comprised three main classes of nasolabial wrinkling: 1, 2, and 3, representing fine, moderate, and deep wrinkles, respectively. A 0 is also used to designate an absence of nasolabial wrinkles. For each main class, a reference photograph was provided as a "gold standard." The nasolabial area was defined as the area between the nasal alar rim and the corner of the mouth. To qualify as a reference photograph, five committee members had to agree on its wrinkle class. To exclude any bias, the photographs presented only the area of the face to be evaluated, rather than the entire face (Figure 1). To take into account possible facial asymmetry, the wrinkle severities of the left and right sides of the face were graded separately. Furthermore, three interclasses could be used to assess wrinkle severity (i.e., 0.5, 1.5, and 2.5) in accordance to the definitions with an estimated wrinkle depth. However, reference photographs were not provided and, thus, these classes were left to the subjective judgment of the assessors. The definitions of the entire classes of the scale are the following:

- Class 0—No wrinkle. No visible wrinkle; continuous skin line.
- Class 0.5—Very shallow yet visible wrinkle.
- Class 1—Fine wrinkle. Visible wrinkle and slight indentation.
- Class 1.5—Visible wrinkle and clear indentation. $<$1-mm wrinkle depth.*
- Class 2—Moderate wrinkle. Clearly visible wrinkle, 1- to 2-mm wrinkle depth.*
- Class 2.5—Prominent and visible wrinkle. More than 2-mm and less than 3-mm wrinkle depth.*

Figure 1. Reference photographs of the four main classes for MFWS and descriptions for all classes.

- Class 3—Deep wrinkle. Deep and furrow wrinkle; more than 3-mm wrinkle depth.*

*Wrinkle depth is based on assessors' estimation rather than physical measurement.

Validation of the MFWS Photographs

The four reference photographs for the MFWS were validated in two stages (Figure 2). The first stage used volunteer photographs. Nine dermatologists or plastic surgeons were initially "trained" with the reference photographs, and then each independently rated an identical set of 40 volunteer photographs showing nasolabial wrinkles of different severity. Assessments were done within 2 hours following training and again 12 to 16 days later. In a second rating session, the photographs were presented in a different order from the first session. The five assessors (of the original nine) who had the highest inter- and intraassessment reliability between the first and second rating systems were selected to continue with the second stage.

The second stage used photographs of clinical study patients. The patients were involved in a clinical study evaluating porcine-derived, collagen-based, injectable filler for the treatment of nasolabial wrinkles. Using the MFWS, each assessor rated the severity of nasolabial folds from identical sets of 100 photographs that displayed right- and left-side frontal views of the nasolabial area. The assessments were done again in a second session 12 to 16 days later.

Descriptions for the MFWS

Descriptions were created to further supplement the reference photographs in the four main classes and to define the interclasses (Figure 1). Three assessors (two dermatologists and one plastic surgeon) were "trained" to grade wrinkle severity by using the reference photographs together with the descriptions. These assessors rated an identical set of 22 volunteer photographs of nasolabial wrinkles. The reliability of this combined approach was then tested using 100 photographs of clinical study patients on two separate sessions separated by an interval of 7 to 13 days.

Statistical Methods

Differences between paired measurements were calculated using descriptive statistics and percentage agreement. However, because some agreement among and within assessors occurred by chance, reliability of the scores was assessed using kappa statistics; Cohen's kappa was used to measure inter- and intraassessor agreement. The kappa coefficient equals 1 if there is perfect agreement, and 0 represents agreement that occurs by chance only.

To determine the level of inter- and intraassessor agreement, a weighted kappa was also calculated that allowed smaller differences between ratings (for example, ratings of "2" and "3") to have a lesser negative impact on the magnitude of the correlation than larger ones (for example, ratings of "0" and "3"). Although there are no absolute cutoffs for kappa coefficients, the kappa interpretation scale of Landis and Koch[18] was applied. Weighted kappa coefficients of >0.61 were regarded as indicating that the MFWS was reliable and ≤ 0.61 as unreliable.

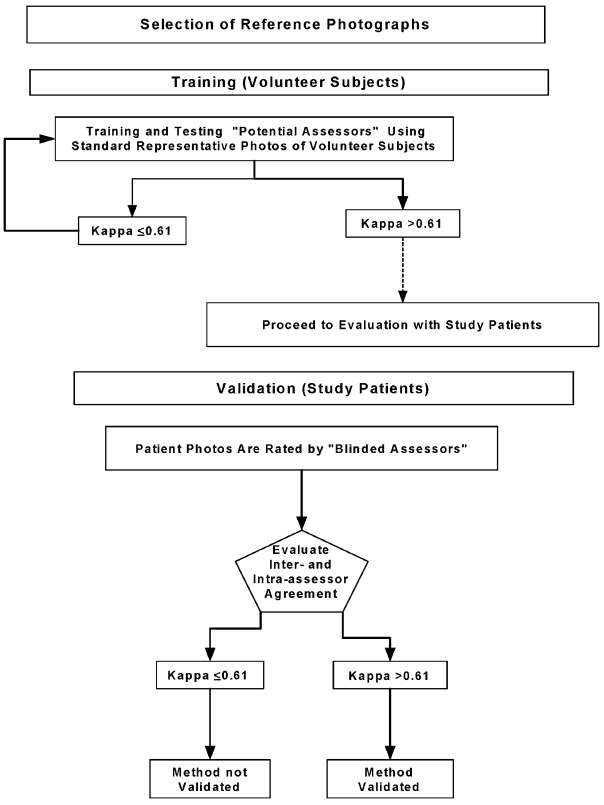

Figure 2. Flow diagram of the validation processes with volunteer and study patients' photographs.

Intraassessor reliability was evaluated by comparison of the test and subsequent retest data for each assessor. Interassessor reliability (internal consistency) was determined by comparing data between pairs of assessors and was expressed as the weighted kappa coefficient for each possible permutation (10 pairs in total). The mean and 95% confidence interval (CI) were calculated for all pairs. The data were analyzed using computer software (SAS, SAS Institute, Cary, NC).

Results

Intraassessor Reliability for Photographic Analysis

In the volunteer photograph study, the weighted kappa for intraassessor reliability was calculated for the nine assessors. Because the data for the left and right side of the face were very similar, the analyses were performed on both sides together (Table 1). The overall weighted kappa was 0.72 (95% CI, 0.68–0.76). Only one assessor was considered unreliable ($\kappa = 0.54$). The five assessors with the highest

TABLE 1. Scaling by Nine Individual Assessors: Values are Kappa Coefficients for Inter- and Intraassessor Reliability in the Study of Volunteer Photographs

	Interassessor		*Intraassessor*
Potential Assessor Number	*Baseline (Day 0)*	*14 ± 2 Days from Baseline*	*14 ± 2 Days from Baseline vs. Baseline*
1	0.70	0.86	0.67
2	0.70	0.82	0.79
3	0.80	0.86	0.79
4	0.84	0.87	0.76
5	0.75	0.67	0.84
6	0.72	0.82	0.54
7	0.87	0.85	0.75
8	0.81	0.74	0.70
9	0.88	0.83	0.68
Overall			0.72 (95% CI, 0.68–0.76)

TABLE 2. Scaling by Five Assessors: Values are Kappa Coefficients for Inter- and Intraassessor Reliability for Clinical Study Patients

	Posttreatment vs. Baseline Visit	
Assessor Number	*Interobserver Coefficient*	*Intraobserver Coefficient*
1	0.68	0.66
2	0.63	0.67
3	0.64	0.70
4	0.69	0.79
5	0.62	0.70

weighted kappa coefficients ($\kappa = 0.75–0.84$) for the retest versus the test visit were then chosen to assess the clinical study patients.

In the study of patients undergoing treatment with porcine-derived, collagen-based injectable filler, intraassessor reliability weighted kappa coefficients for the five assessors for baseline versus posttreatment (12–16 days) are given in Table 2. All assessors had kappa coefficients greater than 0.61 (range, 0.66–0.79). Overall, the coefficients were 0.69 (95% CI, 0.64–0.74) for the left and 0.73 (95% CI, 0.68–0.78) for the right side, with a value of 0.71 (95% CI, 0.68–0.74) for both sides.

Interassessor Reliability for Photographic Analysis

The weighted kappa coefficient for the interassessor reliability was tested for the original nine assessors using the volunteer photographs (Table 1). On the second assessment, the interassessor reliability improved for five of the assessors and was similar for two and worse for two. However, all assessments were reliable within the kappa interpretation scale of Landis and Koch.[18] Four of the five assessors who were chosen for the second stage, which used clinical patient study photographs, had interassessor kappa coefficients greater than 0.8 at the second rating session.

In the clinical patient study, the weighted kappa coefficients for interassessor reliability were 0.66 (95% CI, 0.63–0.69) for the left side, 0.65 (95% CI, 0.63–0.67) for the right side, and 0.65 (95% CI, 0.62–0.68) for both sides. The weighted kappa coefficients ranged from 0.62 to 0.69 (Table 2).

Reliability for Photographic Plus Descriptive Analysis

The results for the weighted kappa coefficients for the intra- and interassessor reliability for the study of the descriptive and visual guidance for the various classes of the MFWS are presented in Table 3. The overall weighted kappa coefficient for the three assessors was 0.79 (95% CI, 0.76–0.82) for the intraassessor reliability and 0.74 (95%CI, 0.69–0.79) for the interassessor reliability.

Discussion

Objective measurements are needed to evaluate the efficacy of antiaging treatments. The MFWS was developed as a simple tool that plastic surgeons and dermatologists could use to assess their treatments.

TABLE 3. Scaling by Three Assessors by Kappa Coefficient (95% CI) Calculated for Inter- and Intraassessor Reliability from Photographs and Definitions of Clinical Study Patients

	Interassessor		*Intraassessor*	
Assessor Number	*Baseline*	*10 ± 3 Days from Baseline*	*Assessor Number*	*10-Day Visit vs. Baseline Visit*
1 vs. 2	0.72 (0.65–0.79)	0.73 (0.66–0.80)	1	0.84 (0.79–0.89)
1 vs. 3	0.79 (0.72–0.85)	0.77 (0.71–0.84)	2	0.71 (0.64–0.78)
2 vs. 3	0.66 (0.57–0.76)	0.71 (0.63–0.78)	3	0.78 (0.72–0.84)
			All	0.79 (0.76–0.82)

In the present validation study, this scale achieved a high level of reliability for the nine assessors who graded the nasolabial folds of volunteers. For the validation process the five assessors with the highest intrarater reliability score had been chosen, a common statistical and procedural practice to minimize the influence of outliers. Reliability was confirmed for the five assessors who graded the clinical study patients injected with filler materials for facial wrinkles and folds. Good statistical inter- and intraassessor agreement indicated that the MFWS grading scale of seven classes was a clinically useful system for the scoring of nasolabial wrinkles. Using photographs alone achieved clinically relevant inter- and intraassessor reliability. However, the addition of definitions to aid the classification by the reference photographs further improved the intra- and interassessor kappa coefficients.

The MFWS has the same number of main classes as the classification[13] from which it was derived. However, clear definitions of the interclasses as well as the three main classes allow easier assessment of the nasolabial fold wrinkle severity than in the original Fitzpatrick classification because of the adaptation of the wrinkle depth in each class to reflect the deeper wrinkling and groove formation typical of the nasolabial fold. Furthermore, the modified grading system evaluates severity of nasolabial fold wrinkling by wrinkle depth, whereas the Fitzpatrick classification is more focused on general wrinkling and elastosis. The modified approach is more relevant for different cosmetic techniques, including injectable fillers or laser treatments, which smooth wrinkle lines and folds and tighten the skin.

Several studies have indicated that analysis of photographs of wrinkles can yield consistent and reliable results. For example, 89.4% of wrinkles were assigned to the same category on a scale of 0 to 5 by eight observers using reference photographs.[8] In this study, the combined use of descriptions and reference photographs to define and grade the wrinkle led to an improvement in wrinkle assessment over the use of photographs alone. This is probably due to the need for a less subjective opinion by the observer regarding the outcome of treatment. These results compare favorably with the 5-point WSRS, which used photographic references and descriptions.[9] In the study using the WSRS, the weighted kappa coefficients for the left (0.77) and right (0.81) sides of the face for intraobserver agreement were similar to those seen in this study.

Therefore, the relatively simple MFWS has proved to be a reliable wrinkle scoring system for nasolabial skin folds. Although sufficiently robust to rely only on four reference photographs, the addition of a series of clear and concise descriptions for each class resulted in greater precision. The MFWS was used in this study to assess wrinkle severity in nasolabial folds but, in addition, it is likely to be adaptable for assessing other skin wrinkles and folds. Training and instruction are needed to ensure proper assessment and grading prior to the first use of the tool by the clinician. In addition, it also has potential for use with equal reproducibility for the evaluation of wrinkles in a clinical setting for live patient evaluation but relies on subjective evaluation by the raters and is not a substitute for other physical methods of measurement. Technological advances (Johnson & Johnson Group of Consumer Companies, Skillman, NJ; Canfield Scientific Inc., Fairfield, NJ) are being made in the three-dimensional volumetric imaging of facial characteristics. These new techniques will undoubtedly increase the reliability of both inter- and intraclinician ratings by finally making it easy to quantify such characteristics as depth of fold at baseline and resultant structural and volumetric changes over time.

Acknowledgments We thank Marina Landau, MD; Klaus Plogmeier, MD; Itzchak Shelkovitz, MD; Rodika Avram, MD; Joseph Ophir, MD; Daphne Thioly Bensoussan, MD; Maximilian Dembinski, MD; Matton Guido, MD; and Birgit Ursula Worle, MD, for their contribution as assessors of this measurement tool validation.

References

1. Koh JS, Kang H, Choi SW, Kim HO. Cigarette smoking associated with premature facial wrinkling: image analysis of facial skin replicas. Int J Dermatol 2002;41:21–7.
2. Gniadecka M. Effects of aging on dermal echogenicity. Skin Res Technol 2001;7:204–7.
3. Lin SJ, Wu R Jr, Tan HY, et al. Evaluating cutaneous photoaging by use of multiphoton fluorescence and second-harmonic generation microscopy. Opt Lett 2005;30:2275–7.
4. Ryu JS, Park SG, Kwak TJ, et al. Improving lip wrinkles: lipstick-related image analysis. Skin Res Technol 2005;11:157–64.
5. 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.
6. Carruthers A, Carruthers J, Said S. Dose-ranging study of botulinum toxin type A in the treatment of glabellar rhytids in females. Dermatol Surg 2005;31:414–22.
7. Wolff EF, Narayan D, Taylor HS. Long-term effects of hormone therapy on skin rigidity and wrinkles. Fertil Steril 2005; 84:285–8.
8. Lemperle G, Holmes RE, Cohen SR, Lemperle SM. A classification of facial wrinkles. Plast Reconstr Surg 2001;108:1735–2001.
9. Day DJ, Littler CM, Swift RW, Gottlieb S. The wrinkle severity rating scale: a validation study. Am J Clin Dermatol 2004;5:49–52.
10. Carruthers A, Carey W, De Lorenzi C, et al. Randomized, double-blind comparison of the efficacy of two hyaluronic acid derivatives, Restylane Perlane and Hylaform, in the treatment of nasolabial folds. Dermatol Surg 2005;31:1591–8.
11. Lindqvist C, Tveten S, Bondevik BE, Fagrell D. A randomized, evaluator-blind, multicenter comparison of the efficacy and tolerability of Perlane versus Zyplast in the correction of nasolabial folds. Plast Reconstr Surg 2005;115:282–9.
12. Stephens T, Tecco M, Menter A. A testing paradigm for evaluating the safety of a cosmetic regimen for sensitive skin individuals [abstract]. J Am Acad Dermatol 2006;54(Suppl. S):40.
13. Fitzpatrick RE, Goldman MP, Satur NM, Tope WD. Pulsed carbon dioxide laser resurfacing of photo-aged facial skin. Arch Dermatol 1996;132:395–402.
14. Weiss RA, Weiss MA, Geronemus RG, McDaniel DH. A novel non-thermal non-ablative full panel LED photomodulation device for reversal of photoaging: digital microscopic and clinical results in various skin types. Drugs Dermatol 2004;3: 605–10.
15. Hall JA, Keller PJ, Keller GS. Dose response of combination photorejuvenation using intense pulsed light-activated photodynamic therapy and radiofrequency energy. Arch Facial Plast Surg 2004;6:374–8.
16. Kopera D, Smolle J, Kaddu S, Kerl H. Nonablative laser treatment of wrinkles: meeting the objective? Assessment by 25 dermatologists. Br J Dermatol 2004;150:936–9.
17. Monstrey SJ, Pitaru S, Hamdi M, et al. A two-stage phase I trial of EVOLENCE™ 30 collagen for soft-tissue contour correction. Plast Reconstr Surg 2007;120:303–11.
18. Landis JR, Koch GC. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics 1977;33:363–74.

Address correspondence and reprint requests to: David Shoshani, MD, ColBar LifeScience Ltd, 9, Hamenofim Street, Tower A, PO Box 12206, Herzliya 46733, Israel, or e-mail: davids@colbar.com

COMMENTARY

I agree with the comments of Dr. Shoshani and colleagues that although there are many wrinkle and photodamage scales in use, it is confusing to choose a reliable scale. When we were working on CO_2 laser resurfacing, I thought that it was important to be able to grade changes in skin and wrinkles in the most objective way that we could. There were no scales that used reference photos at that time, and the other scales included dyschromia, telangiectasia, skin cancer, and actinic keratoses as well. I developed the wrinkle scale as a means of evaluating the degree of improvement in texture and lines, i.e., secondary to new collagen formation. New collagen formation was the most significant change induced by CO_2 resurfacing, and our scale of 3 classes defined by verbal description, but referenced by 3 photos in each class, has been adopted by the FDA as the standard for measuring improvement in texture and lines.

Dr. Shoshani and colleagues are to be congratulated for adapting the scale for the use in specifically addressing the evaluation of improvement in the nasolabial fold, as this deep line or fold was not a focus of the original scale. In order to evaluate volume and line changes in the nasolabial fold, its own reference photos are needed. Their scale has been validated and should prove be very useful.

RICHARD E. FITZPATRICK, MD
Carlsbad, CA