Accepted: 10 September 2017

DOI: 10.1111/srt.12413

ORIGINAL ARTICLE

WILEY

Validation of digital photographic reference scales for evaluating facial aging signs

Randa Jdid[1] | Julie Latreille[1] | Frédérique Soppelsa[2] | Erwin Tschachler[3,4] | Frédérique Morizot[1]

[1]Department of Biology & Women's Beauty, Chanel, Pantin, France

[2]Department of Biometrics, Lincoln, Boulogne Billancourt, France

[3]CE.R.I.E.S., Neuilly sur Seine, France

[4]Department of Dermatology, Medical University of Vienna, Vienna, Austria

Correspondence
Randa Jdid, 8 rue du Cheval Blanc, F - 93694 Pantin Cedex, France.
Email: randa.jdid@chanel-corp.com

Abstract

Background: Validated tools are essential to evaluate facial skin aging for both dermatological and cosmetic investigations. While many visual aging scales have been developed, few have been validated and none in terms of degree of distinguishability (DD). We developed and validated a series of visual scales using a novel digital interface for scoring facial skin aging in Caucasian women.

Materials and methods: Three dermatologists independently established scales for 12 distinct aging signs from high-definition facial photographs of 400 adult women (Fitzpatrick phototypes I-IV) taken under standardized conditions. They then selected a consensus scale for each individual sign with a representative photo per grade. Scales were integrated into a digital interface allowing simultaneous viewing of all grades of each scale alongside the photograph of a test subject. Next, scales were validated by a different dermatologist, a general practitioner and a non-medical expert skin evaluator using photos of 350 women which had not been used for establishing the scales.

Results: Kappa estimates showed almost perfect agreement for wrinkle and skin aging scales (≥0.85) and moderate to substantial agreement for scales relating to color irregularities (telangiectasia, solar lentigines, freckles) for both inter- and intra-observer reproducibility. Intra-observer DD estimates were mostly high. Non-dermatologists performed well on reproducibility for both Kappa (from 0.6 to 0.9) and DD estimates.

Conclusion: Our work demonstrates that the digital interface scales for 12 distinct aging features are highly suitable for use in clinical and epidemiological studies on skin aging by both dermatologists and non-dermatologists.

KEYWORDS

degree of distinguishability, digital interface, facial aging, validated skin scale

1 | INTRODUCTION

The interest in objectively studying aging-related changes of the human face has seen a dramatic rise over the past years. An impressive surge of surgical and non-surgical "anti-aging" procedures over the past decades[1] has resulted in an increased demand for tools to objectively assess the baseline degree of aging and the post-interventional therapeutic success. Apart from the relevance of such investigations for esthetic interventions, there has also been a surge of interest in pinpointing genetic factors contributing to accelerated or slowed-down aging.[2-6] Precise validated scales for the different signs of skin aging are pivotal tools to advance this field of research.

The interest in using scales based on photographic images for the evaluation of cutaneous photodamage, as well as dermatological and

cosmetic treatments, was acknowledged by Griffiths et al[7] who reported that photographic scales offered a much more reliable means of evaluating photodamage than purely descriptive ordinal scales. Since then, a number of printed photographic scales for a range of skin aging signs and for different skin types have been published.[8-11] However, while many different photographic scales exist today, most of them have not been validated and/or are not available for general use.[1,8-11]

For the establishment of visual aging scales, two different approaches have been taken in the past: Global scales and scales for specific aging signs. While global aging scales can be useful in clinical and epidemiological trials to summarize given phenomena,[12] scales evaluating individual aging sign have the advantage that they allow for the collection of more precise and detailed information.[13-18] In addition, they can be applied to population with different ethnic background. For this reason, we focused our present efforts on the establishment and validation of scales of the latter type.

The present study was designed to develop a series of extended scales for 12 distinct signs related to facial aging. We took advantage of the recent progress made in the field of image acquisition technology and viewing softwares[12] to integrate these scales into an information technology interface to facilitate scoring of facial images to be studied. Each scale was validated and refined with two statistical approaches (1) inter and intra-observer agreement using the Kappa coefficient and (2) a more recently developed novel statistical approach which evaluates the degree of distinguishability (DD) between two consecutive grades, allowing the equivalence of "distances" between two grades to be determined for the scale as a whole.[19]

2 | METHODS

2.1 | Subjects

Photographs of 350 Caucasian women from six age groups (20-80 years old) were used. At the time of inclusion in the photographic database, women were ineligible if they had visible or permanent makeup, adornments, colored contact lenses, uniquely identifiable characteristics (eg, scars, birthmarks, tattoos), skin abnormalities or had undergone esthetic procedures. Written informed consent was obtained from all subjects at the time the photographs were taken allowing their use for research studies.

2.2 | Study design

The study was performed in two steps using digital photographs of Caucasian women (Fitzpatrick phototypes I-IV). Initially, a series of photographic scales for 12 signs of facial aging with up to nine grades per scale were established by three dermatologists from a database of 400 digital facial images, then integrated into an electronic interface using specifically designed software. The scales were subsequently validated in terms of reproducibility and degree of distinguishability between grades, by three independent evaluators of different dermatological backgrounds using an independent collection of facial photographs. Different subject cohorts were used for establishing and validating the scales. Scales were then adapted taking into account statistically identified weaknesses, by combining grades or changing the representative image of a given grade to obtain final consensus scales. The study was performed in conformance with the Declaration of Helsinki.

2.3 | Photographs

A closed photographic system was used to ensure accurate and reproducible subject positioning and controlled lighting. Each face was illuminated by three flashes: one in front of the face (diffuse light), the height of this flash was adjusted to the height of the subject's face; and two flashes illuminating the face from a 45° angle (direct light), the height of these flashes was fixed. These lighting conditions were defined to avoid cast shadows and minimize variation from shading on the faces. Four standardized, high-resolution digital images of the face were taken for each participant (one frontal view with open and closed eye and one of each profile) by a trained technician using a Canon EOS-1 Ds Mark II, 17 MP camera after subjects had rested in an air-conditioned room for 30 minutes. A chart containing 48 color patches on each picture was used to allow for color calibration. Hair was covered by a headband and a series of photos was taken, face-on (full face) and profile, with eyes open and closed. Only good quality photos of subjects with neutral facial expression (no smile, frown, visible teeth, partially open eyes, etc.) were used. Photos were presented at a uniform magnification and size.

2.4 | Development of photographic scales for facial aging signs

Separate scales were developed for 12 signs of facial aging (Table 1); solar lentigines "age spots" (face-on forehead, profile cheek), freckles (face-on forehead, profile cheek), expression lines (horizontal; face-on forehead), frown lines (vertical; face-on forehead), wrinkles under the eyes (face-on of the open eye), drooping eyelid (face-on closed eye), wrinkles on the upper lip (face-on), marionette lines (face-on), telangiectasia (profile), crow's feet wrinkles (profile), nasolabial fold (profile), and loss of facial oval (profile).

For each aging sign, 40-50 photos illustrating the full range of severity were selected by a dermatologist from a panel of photographs of 400 women, using existing scales as a guide to determine the range of grades.[13] Photos were cropped, maintaining proportionality, to present only the relevant aging sign and printed in triplicate. Three additional dermatologists independently developed a scale of up to 9 grades for each sign then agreed on a single consensus scale per sign with a representative photo per grade, from 0 (no sign of aging) to the highest severity of aging, and were designed to have an even difference between grades. The total number of grades varied depending on the aging sign.

TABLE 1 Inter-observer agreement between grades assigned by the three independent evaluators for facial aging signs in 350 subjects, Evaluation 1

Aging sign	Scale grade range	Kappa coefficient for paired evaluators		
		Dermatologist/GP	Dermatologist/Skin expert	GP/Skin expert
Frown lines	0-8	0.88	0.89	0.92
Expression lines	0-7	0.87	0.86	0.90
Crow's feet wrinkles	0-9	0.90	0.92	0.93
Wrinkles under eyes	0-8	0.83	0.80	0.81
Wrinkles upper lip	0-8	0.88	0.88	0.88
Marionette lines	0-6	0.83	0.75	0.83
Drooping eyelid	0-7	0.88	0.84	0.85
Nasolabial fold	0-8	0.89	0.87	0.89
Loss of facial oval	0-8	0.85	0.88	0.86
Telangiectasia	0-6	0.75	0.65	0.70
Freckles (F)	0-4	0.52	0.44	0.75
Freckles (C)	0-4	0.52	0.39	0.67
Solar lentigines (F)	0-6	0.75	0.64	0.72
Solar lentigines (C)	0-6	0.82	0.66	0.75

C, cheek; F, forehead; GP, general practitioner.
Landis & Koch: 0.2 represents slight agreement; >0.2 to 0.4 fair agreement; >0.4 to 0.6 moderate agreement; >0.6 to 0.8 substantial agreement; and >0.8 to 1 almost perfect agreement.

2.5 | Software for photographic scales

Consensus scale photos were integrated into the Digital Viewing Interface developed by QuantifiCare, S.A. (Sophia Antipolis, France). All photos from a given scale were visible on the screen alongside the respective facial image of the subject to be evaluated. The four digital standardized photos were available for each test subject, with two standardized enlargements of each image (Fig. 1).

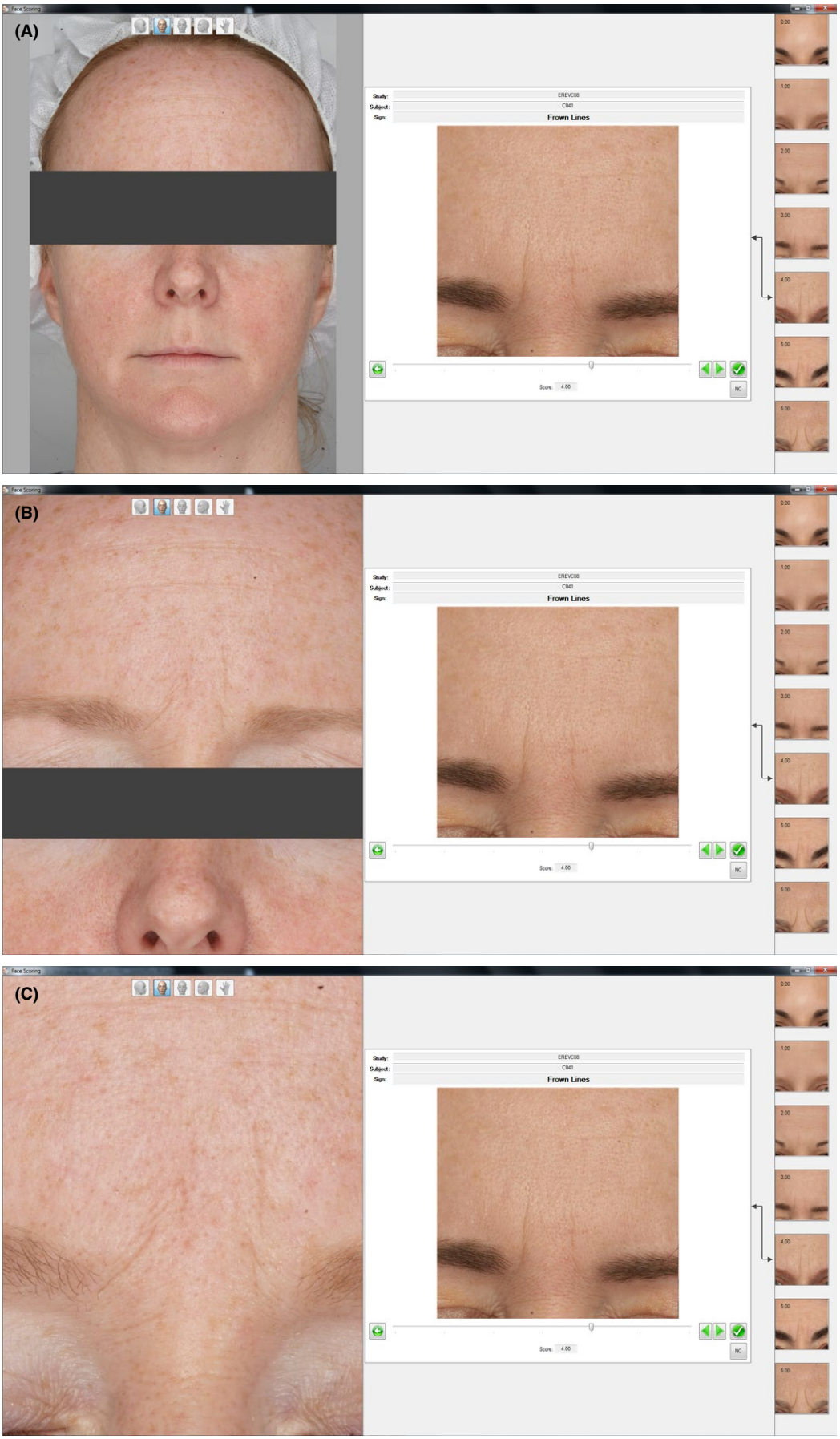
2.6 | Validation of photographic aging scales

Validation was carried out independently by three evaluators, a dermatologist (not from the scale development panel), a general practitioner and a non-medical expert skin evaluator (pharmacist) using a panel of photographs of 374 women which had not been used for the establishment of the scales. The evaluators were first trained in the use of the software by scoring photographs of 24 women. Each evaluator then scored photographs from the remaining 350 women for each of the 12 scales over a period of 18 half-days (~20 photos per half-day). The evaluation was repeated after 1 month.

2.7 | Statistical considerations

Photo presentation for scoring during the validation process was randomized with stratification based on age. The same randomization series was presented to each evaluator for all scales. Reproducibility for each scale was assessed using two methods. (1) Intra-observer agreement and inter-observer repeatability were quantified using the weighted Kappa coefficient.[20] Outcomes were interpreted according to Landis and Koch as follows; <0.2, representing slight agreement, >0.2 to 0.4 fair agreement, >0.4 to 0.6 moderate agreement, >0.6 to 0.8 substantial agreement, and >0.8 to 1 almost perfect agreement.[21] (2) The degree of the observer's DD between adjacent categories in each scale and the homogeneity between these differences was estimated using a log-linear non-uniform association model.[19] Values range from 0 (not distinguishable) to 1 (perfectly distinguishable). DD values ≥0.6 represent good distinguishability between grades. When more than one observer had difficulty distinguishing between two grades, only one of the two grades was kept in the scale or the representative photo for one of the two grades was replaced. Kappa and DD estimates were expressed with their 95% confidence intervals (CI) where appropriate. Analyses were performed using SAS software (version 9.1.3; SAS Institute, Cary, NC, USA) and R software (version 2.15.2).[22]

FIGURE 1 Digital viewing interface for evaluating facial aging signs using a photographic scale. Example screen shots of the photo imaging software for frown lines scale showing the subject to be evaluated on the left (face-on or profile views can be selected from the four miniature images at the top of main subject image). Each grade of the scale is shown on the right and the grade selected is shown on the middle. Images of the subject to be evaluated were available at a standard size (A) and two standardized enlargements (B and C). Eyes are masked for the purposes of publication only


(A)
Face Scoring
Study:
Subject:
Sign:
Frown Lines
Score: 4.00
NC
(B)
Face Scoring
Study:
Subject:
Sign:
Frown Lines
Score: 4.00
NC
(C)
Face Scoring
Study:
Subject:
Sign:
Frown Lines
Score: 4.00
NC

3 | RESULTS

3.1 | Reference scales for facial aging signs

Consensus scales were established for each of the 12 distinct aging signs (Table 1) in terms of the number of grades and the representative photo per grade as described in material and methods. The number of grades per scale ranged from 0 to 4 through 0 to 9, depending on the sign (Table 1). Representative screen shots of the digital interface used to visualize a scale and evaluate subjects are shown in Figure 1.

3.2 | Inter-observer agreement and intra-observer reproducibility

Reproducibility of each scale was evaluated for each sign using photos from 350 women who were different from the ones used for the establishment of the scales; 20-29 years (N = 37), 30-39 years (N = 59), 40-49 years (N = 77), 50-59 years (N = 74), 60-69 years (N = 71), ≥70 years (N = 32). The number of women included below the age of 30 years and above 70 years was lower in order to minimize skewing the outcome, since scoring of aging signs in these age groups is easier, making higher agreement more likely. The degree of inter-observer agreement according to Kappa estimates is shown in Tables 1 and 2 for each aging sign for the two consecutive evaluations performed 1 month apart. For both evaluations, agreement between each pair of evaluators was very high (>0.8) for the majority of wrinkle and sagging signs. Agreement was mostly substantial (>0.6-0.8) for evaluation of telangiectasia and lentigines, and moderate to substantial (>0.4-0.6) for freckles.

A similar profile was seen for intra-observer agreement between the first and second evaluations, with almost perfect agreement for wrinkles and sagging signs, substantial to almost perfect agreement for telangiectasia and solar lentigines, and substantial agreement for freckles (Table 3). Kappa values for intra-observer agreement were similar for each sign between all three evaluators.

3.3 | Degree of distinguishability and scale adjustment

For all scales other than expression lines, intra-observer DD evaluations showed reasonably good agreement for most adjacent grades. When poor agreement was seen, scales were optimized by the suppression of one grade or by replacing a photo for a given grade (DD values <0.6). Difficulties distinguishing between two consecutive grades were identified for "wrinkles under the eyes," "crow's feet," "wrinkles on the upper lip," "telangiectasia," "nasolabial fold", and "age spots." All these scales were refined by removal of a grade for each, while two of the initial 10 grades were removed for "loss of facial oval shape" and "frown lines." For drooping eyelids, the photo for one grade was replaced. The cheek and forehead freckle scales were replaced by a single freckle scale of the whole face.

Heterogeneity in intra-observer DD estimates due to sporadic scoring difficulties as reflected by wide 95% CIs, was seen for all evaluators for most grades, notably for most wrinkles (crow's feet, upper lip, under the eye and frown lines) and sagging but to a lesser extent for expression and marionette lines and the three pigmentation scales.

TABLE 2 Inter-observer agreement between grades assigned by the three independent evaluators for facial aging signs in 350 subjects, Evaluation 2[a]

		Kappa coefficient for paired evaluators		
Aging sign	Scale grade range	Dermatologist/GP	Dermatologist/Skin expert	GP/Skin expert
Frown lines	0-8	0.85	0.85	0.91
Expression lines	0-7	0.84	0.79	0.87
Crow's feet wrinkles	0-9	0.87	0.89	0.92
Wrinkles under eyes	0-8	0.85	0.82	0.83
Wrinkles upper lip	0-8	0.91	0.86	0.90
Marionette lines	0-6	0.80	0.80	0.90
Drooping eyelid	0-7	0.91	0.79	0.80
Nasolabial fold	0-8	0.90	0.86	0.85
Loss of facial oval	0-8	0.90	0.87	0.89
Telangiectasia	0-6	0.52	0.69	0.65
Freckles (F)	0-4	0.57	0.62	0.66
Freckles (C)	0-4	0.55	0.45	0.59
Solar lentigines (F)	0-6	0.68	0.68	0.72
Solar lentigines (C)	0-6	0.79	0.75	0.71

C, cheek; F, forehead; GP, general practitioner.
Landis & Koch: 0.2 represents slight agreement; >0.2 to 0.4 fair agreement; >0.4 to 0.6 moderate agreement; >0.6 to 0.8 substantial agreement; and >0.8 to 1 almost perfect agreement.
[a]Performed 1 month after Evaluation 1.

TABLE 3 Intra-observer agreement between grades assigned by the three evaluators for facial aging signs in 350 subjects

		Kappa coefficient for each evaluator		
Aging sign	**Scale grade range**	**Dermatologist**	**GP**	**Skin expert**
Frown lines	0-8	0.89	0.94	0.93
Expression lines	0-7	0.87	0.91	0.92
Crow's feet wrinkles	0-9	0.88	0.94	0.94
Wrinkles under eyes	0-8	0.84	0.89	0.86
Wrinkles upper lip	0-8	0.92	0.91	0.88
Marionette lines	0-6	0.90	0.89	0.90
Drooping eyelid	0-7	0.91	0.91	0.90
Nasolabial fold	0-8	0.91	0.91	0.89
Loss of facial oval	0-8	0.88	0.90	0.93
Telangiectasia	0-6	0.80	0.81	0.73
Freckles (F)	0-4	0.60	0.74	0.65
Freckles (C)	0-4	0.67	0.71	0.62
Solar lentigines (F)	0-6	0.77	0.79	0.64
Solar lentigines (C)	0-6	0.84	0.84	0.65

C, cheek; F, forehead; GP, general practitioner.
Landis & Koch: 0.2 represents slight agreement; >0.2 to 0.4 fair agreement; >0.4 to 0.6 moderate agreement; >0.6 to 0.8 substantial agreement; and >0.8 to 1 almost perfect agreement.

Inter-observer heterogeneity was also seen between evaluators in several scales.

4 | DISCUSSION

In this report, we developed a series of scales for 12 distinct aging signs with equivalence of distance between grades for evaluating facial aging in Caucasian women, and optimized their use by integrating them into a novel Digital Viewing Interface. Reproducibility in 350 subjects in terms of inter-observer and intra-observer agreement was high overall with most Kappa estimates being 0.85 or higher for all wrinkle and skin sagging scales, reflecting almost perfect agreement. Assessments using scales for pigmentation disorders (freckles, age spots) and telangiectasia were less reproducible than with the scales for wrinkles and sagging. This may reflect the wide variation in size, color, skin contrast and distribution of color irregularities and the resulting difficulty in developing continuous scales.[13] Overall, improvements of the present scales as compared to already existing ones, concern the large number of images used as starting material to build the scales as well as the quality of the images. In addition, the incorporation into a straightforward digital interface makes the scales easy to use also for unexperienced investigators.

Evaluation of distinguishability provides valuable information on the structure of a scale.[23] This method adds another dimension to testing the validity of ordinal scales by highlighting more precisely where they can be improved.[24] Despite strong intra- and inter-observer agreement, DD analyses identified limiting weaknesses in the initial steps of building the present scales by identifying the grades in given scales which did not allow for reproducible scoring. This approach permitted us to refine the scales by either pooling adjacent grades or replacing representative photos to improve the accuracy. Indeed, an equal and high DD value between all adjacent categories of the scale would clearly lead to a lower variability within ratings and hence a better agreement between them.[23]

The digital interface used in this study, is a practical and rapid means for simultaneously visualizing the entire scale and multiple angles of the image of the study subject. It also allows the user to switch rapidly between enlargements of images under evaluation. Furthermore, it minimizes the potential for errors and allows direct data capture of the analysis database. The excellent levels of reproducibility reported in this study support the use of the digital interface as an adapted tool for scoring aging signs.

In our study, also non-dermatologist examiners demonstrated good intra-observer reproducibility, which is likely a reflection of the adequacy of the instructions provided in the training and the straightforwardness of the digital interphase. Its usefulness for non-dermatological specialists highlights the potential for their wide application.

In conclusion, we have developed new validated scales for 12 signs of facial aging with equivalence of distance between the grades, available in a digital interface allowing simultaneous visualization of the scales alongside the facial image to be scored. These scales are suitable for both clinical studies evaluating the effect of treatments on skin aging as well as for epidemiological studies comparing the rate of aging in different individuals of large cohorts. The scales and the digital interface are made available on the Quantificare website https://cloud.quantificare.com/s/9VBBpxTQv74VqaF.

ACKNOWLEDGEMENTS

All authors declare no conflicts of interest. We thank Pr Barbara Gilchrest for critical readings of the manuscript, the investigators of

EVIC FRANCE involved in this study, and Dr Sarah MacKenzie (Medi. Axe, France) for medical writing assistance.

ORCID

Randa Jdid http://orcid.org/0000-0002-0465-2474

REFERENCES

1. Carruthers A, Carruthers J. A validated facial grading scale: the future of facial ageing measurement tools? *J Cosmet Laser Ther*. 2010;12:235-241.
2. Le Clerc S, Taing L, Ezzedine K, et al. A genome-wide association study in Caucasian women points out a putative role of the STXBP5L gene in facial photoaging. *J Invest Dermatol*. 2013;133:929-935.
3. Chang ALS, Atzmon G, Bergman A, et al. Identification of genes promoting skin youthfulness by genome-wide association study. *J Invest Dermatol*. 2014;134:651-657.
4. Zhang M, Qureshi AA, Hunter DJ, et al. A genome-wide association study of severe teenage acne in European Americans. *Hum Genet*. 2014;133:259-264.
5. Navarini AA, Simpson MA, Weale M, et al. Genome-wide association study identifies three novel susceptibility loci for severe Acne vulgaris. *Nat Commun*. 2014;5:4020.
6. Jacobs LC, Hamer MA, Gunn DA, et al. A genome-wide association study identifies the skin color genes IRF4, MC1R, ASIP, and BNC2 influencing facial pigmented spots. *J Invest Dermatol*. 2015;135:1735-1742.
7. Griffiths CE, Wang TS, Hamilton TA, et al. A photonumeric scale for the assessment of cutaneous photodamage. *Arch Dermatol*. 1992;128:347-351.
8. Bazin R. *Skin aging atlas. Volume 1, Caucasian type*. Paris: Med'com, 2007.
9. Bazin R, Flament F. *Skin aging atlas. Volume 2, Asian type*. Paris: Med'com, 2010.
10. Bazin R, Flament F, Giron F. *Skin aging atlas. Volume 3, Afro-American type*. Paris: Med'com, 2012.
11. Bernois A, Huber A, Derome C, et al. A photographic scale for the evaluation of facial skin aging in Indian women. *Eur J Dermatol*. 2011;21:700-704.
12. Larnier C, Ortonne JP, Venot A, et al. Evaluation of cutaneous photodamage using a photographic scale. *Br J Dermatol*. 1994;130:167-173.
13. Morizot F, Lopez C, Guinot M et al. Development of photographic scales documenting features of skin ageing based on digital images. 20th World Congress of Dermatology [Abstract]. *Ann Dermatol Venereol* 2002: 129: 1s402.
14. Day DJ, Littler CM, Swift RW, et al. The wrinkle severity rating scale: a validation study. *Am J Clin Dermatol*. 2004;5:49-52.
15. Flynn TC, Carruthers A, Carruthers J, et al. Validated assessment scales for the upper face. *Dermatol Surg*. 2012;38:309-319.
16. Carruthers A, Carruthers J, Hardas B, et al. A validated grading scale for marionette lines. *Dermatol Surg*. 2008;34(Suppl 2):S167-172.
17. Carruthers A, Carruthers J, Hardas B, et al. A validated grading scale for forehead lines. *Dermatol Surg*. 2008;34(Suppl 2):S155-160.
18. Carruthers A, Carruthers J, Hardas B, et al. A validated grading scale for crow's feet. *Dermatol Surg*. 2008;34(Suppl 2):S173-178.
19. Valet F, Guinot C, Mary JY. Log-linear non-uniform association models for agreement between two ratings on an ordinal scale. *Stat Med*. 2007;26:647-662.
20. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull*. 1968;70:213-220.
21. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159-174.
22. R Core Team (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/ [Internet]. https://doi.org/www.r-project.org/
23. Valet F, Guinot C, Ezzedine K, et al. Quality assessment of ordinal scale reproducibility: log-linear models provided useful information on scale structure. *J Clin Epidemiol*. 2008;61:983-990.
24. Valet F, Ezzedine K, Malvy D, et al. Assessing the reliability of four severity scales depicting skin ageing features. *Br J Dermatol*. 2009;161:153-158.

How to cite this article: Jdid R, Latreille J, Soppelsa F, Tschachler E, Morizot F. Validation of digital photographic reference scales for evaluating facial aging signs. *Skin Res Technol*. 2018;24:196–202. https://doi.org/10.1111/srt.12413