Validation of image analysis techniques to measure skin aging features from facial photographs

M.A. Hamer
L.C. Jacobs
J.S. Lall
A. Wollstein
L.M. Hollestein
A.R. Rae
K.W. Gossage
A. Hofman
F. Liu
M. Kayser
T. Nijsten
D.A. Gunn

Skin Res Technol. 2015 Nov;21(4):392-402

ABSTRACT

Background: Accurate measurement of the extent of skin aging is challenging, but crucial for research. Image analysis offers a quick and consistent approach for quantifying skin aging features from photographs, but is prone to technical bias and requires proper validation.

Methods: Facial photographs of 75 male and 75 female northwestern European participants, randomly selected from the Rotterdam Study, were graded by two physicians using photonumeric scales for wrinkles (full face, forehead, crow's feet, nasolabial fold and upper lip), pigmented spots and telangiectasia. Image analysis measurements of the same features were optimized using photonumeric grades from 50 participants, then compared to photonumeric grading in the 100 remaining participants stratified by sex.

Results: The inter-rater reliability of the photonumeric grades was good to excellent (intraclass correlation coefficients 0.65-0.93). Correlations between the digital measures and the photonumeric grading were moderate to excellent for all the wrinkle comparisons (Spearman's rho ρ=0.52-0.89) bar the upper lip wrinkles in the men (fair, ρ=0.30). Correlations were moderate to good for pigmented spots and telangiectasia (ρ=0.60-0.75).

Conclusion: These comparisons demonstrate that all the image analysis measures, bar the upper lip measure in the men, are suitable for use in skin aging research and highlight areas of improvement for future refinements of the techniques.

INTRODUCTION

Skin aging is a heterogeneous phenotype, which includes features such as wrinkles, pigmented spots, and telangiectasia (i.e. red veins). During the last few decades, people have become increasingly concerned about their appearance, with facial skin aging being a critical component[1]. Consequently, basic and clinical research on this topic has expanded rapidly. To measure the degree that skin has visibly aged, several different photonumeric scales have been published, which are feature specific or a combination of different skin aging features[2-5]. However, a recognized gold standard scale for skin aging is still lacking.

Griffiths et al[2] introduced one of the first facial skin aging scales, assessing photoaging as a single entity, combining wrinkles, pigmented spots and telangiectasia in a 9-point scale. Larnier et al[3] also created a photonumeric scale, but introduced three different photographs per grade to cover the variable nature of photodamage. Subsequently, photonumeric scales for wrinkles at different facial sites were created to evaluate aesthetic procedures, either using photographs[4] or computer-simulated images[6]. Other scales differentiated between the relative contribution of intrinsic vs. extrinsic factors to facial skin aging[7,8]. For pigmented spots, a few photonumeric severity scales are available for Caucasian[9-11] and non-Caucasian populations[12]. For telangiectasia, available scales mainly capture improvement after cosmetic procedures[13]. Only a few scales have been published for epidemiological purposes, either descriptive[8,14] or photonumeric[11]. However, the inter-observer agreement for the photonumeric scale was rather low and only telangiectasia in the crow's feet area were taken into account[11].

In addition to these categorical scales, there are quantitative rating scales that measure three-dimensional (3D) details of the skin surface using skin replicas[4,15] or computer-assisted skin surface topography[16]. Raking light optical profilometry applied directly to facial photography[17] is another method to quantitatively measure wrinkles, providing multiple wrinkle parameters, including wrinkle number, length, width, area and depth. Correlations with photonumeric grading of crow's feet were good, although correlations for the other facial sites were not mentioned[17]. Recently, a 3D fringe projection method was used to measure facial wrinkles[18,19]. It was utilized to estimate the likelihood of the lifetime development of wrinkles, based on wrinkle differences between age groups[18]. Digital measures previously developed for pigmented spots measure the affected skin area using various image analysis techniques[20-22]. However, none of these techniques, nor image analysis techniques for measuring telangiectasia, have been validated against photonumeric grading.

The potential advantages of digital measurements are their sensitivity, reliability and generation of continuous outcomes. In contrast to digital methods, photonumeric grading can be unwittingly influenced by other features of aging such as hair graying or facial sagging. In addition, it seems plausible although speculative that digital measurement is more sensitive to subtle pre-clinical aging, which is not always visible to the human eye. Digital measurement is also time-saving which is of benefit for research, particularly in large cohorts. Furthermore, a continuous digital measure

may detect smaller differences between individuals and, therefore, have more power to detect associations in observational studies compared to photonumeric categorical scales. However, technical influences (e.g. variations in lighting) affect image analysis techniques and hence blinded tests are required to determine the similarity of the digital measures with human expert assessment.

The aim of this study is to create and validate digital measurements for wrinkles, pigmented spots and telangiectasia, using high-resolution digital photographs.

METHODS

Study population

The Rotterdam Study (RS) is a prospective population-based cohort study conducted in Ommoord, a suburb of Rotterdam, the Netherlands. Details of the study design and objectives have been described elsewhere[23]. From August 2010 onwards, standardized high-resolution digital 3D facial photographs were collected on participants at the RS center (N=4648 to date). The current study included images of 150 participants, all of northwestern European ancestry. The RS has been approved by the medical ethics committee according to the Wet Bevolkingsonderzoek ERGO (Population Study Act Rotterdam Study), executed by the Ministry of Health, Welfare and Sports of the Netherlands and all participants provided written informed consent.

Image acquisition

For all participants, high resolution standardized full face photographs were obtained with a Premier 3dMD face3-plus UHD camera (3dMD, Atlanta, GA, USA), in a room without daylight. Participants focused on a standardized viewpoint and were asked not to wear any make-up, facial cream, or jewelry. Three two-dimensional (2D) photographs (2452 × 2056 pixels, 14.7MB in BMP format) were taken simultaneously from three prefixed angles (one upper frontal and two 45° lateral photos). By combining these photos, the 3dMD software (www.3dmd.com) created an image containing 3D information of the whole face. The machine was calibrated daily to control for camera position and environmental light intensity.

Photonumeric grading

We created new 5-point scales for full face wrinkles, pigmented spots and telangiectasia. Full facial wrinkles have different patterns in men and women[18,24,25] but there are no sex-specific scales available in the literature. Therefore, we established new sex-specific scales for full face wrinkling, based on photodamage grading scales by Griffiths et al.[2] and Larnier et al.[3], using images from the RS. For pigmented spots and telangiectasia, there was no accessible photonumeric scale available beyond the crow's feet area. Therefore, we created new scales as for global wrinkles, but these were not sex-specific because there seemed to be little difference in facial location of pigmented spots and telangiectasia between men and women. Pigmented spots included both solar lentigines

and seborrheic keratoses. Freckles, nevi and actinic keratoses were not considered as pigmented spots. For telangiectasia grading, we took into account only red and purple-blue vein like structures as well as spider nevi. Erythema, red papules and other reddish structures in the face were ignored.

For the photonumeric grading of the forehead, crow's feet, nasolabial fold and upper lip wrinkles, we used the Skin Aging Atlas book[26], which is based on several published scales[2-4,27]. The scales within the book are focused solely on the depth of the deepest wrinkle but for the crow's feet area, a scale for the number of wrinkles was also available. Hence, for the crow's feet we also generated an overall wrinkle severity score ((number + depth)/2). The location-specific scales consisted of either 6 or 7 grades[26]. In order to create uniformity, we only used six grades and in case of seven, we discarded the lowest one, considering that our study was conducted in an elderly population.

For all skin aging features (full face wrinkles, forehead wrinkles, crow's feet, nasolabial fold, upper lip wrinkles, pigmented spots and telangiectasia), an optimization set of 50 photos was graded by two independent physicians (MAH and LCJ) for all seven features. Subsequently the two physicians discussed any grading differences and reached a consensus grade; these grades were also used to optimize the digital measurements. A validation set of 100 photos was then graded blindly by the two physicians for the same seven features.

Masking of photographs

Full face wrinkles

For quantification of wrinkles on the whole face, standardized 2D front and side images were generated from the 3D rendering (1920×1080 pixels, 1MB in TIF format) using Blender (http://www.blender.org/v2.7) as the original front 2D photographs were taken from above the participants, causing the chin to be tilted away from the camera reducing the area of skin visible. The photographs were masked to isolate the skin areas in the image using semi-automated masking (MATLAB, The MathWorks, Inc, Natick, MA, USA, version 2013a), Figure 1A.

Wrinkles per localized facial site

The original 2D photographs of the left-hand side of the face were used to measure wrinkle severity at localized facial sites as they had a higher resolution than their 3D equivalent. A bespoke semi-automated program cropped localized sites (forehead, crow's feet, nasolabial fold and upper lip) from each image, Figure 1B.

Pigmented spots

The 2D front photographs were used to generate the pigmented spots digital measure, since the higher resolution was necessary to detect subtle color differences of the skin between small objects (e.g. pores versus senile lentigines). Masking was applied to each image similar to the full face wrinkle masking but additionally excluding the jaw and mouth area (Figure 1C), because stubble in men can influence the measurement.

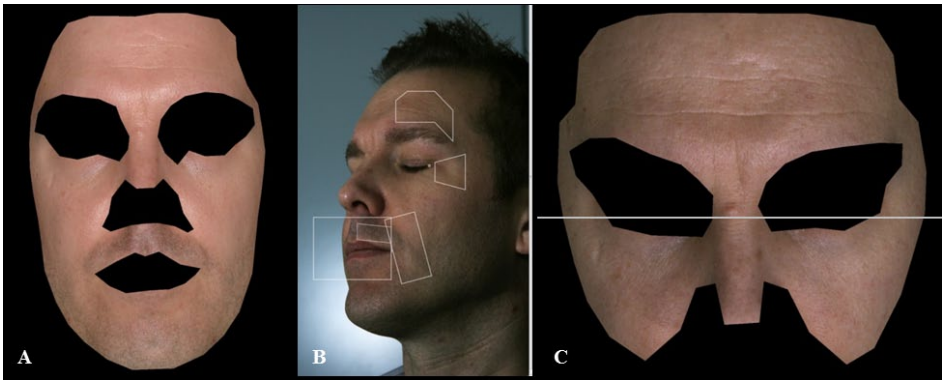

Figure 1. Examples of masking and the delineation of localized sites in images. (**A**) Masking of an image produced from the 3D rendering for full face wrinkle measurement. Non-skin features that could be detected as wrinkles (i.e. eyes, eyebrows, hair, ears, nostrils, and lips) were masked as well as the shadowing that was present along outer most lateral sides of the face. A mask was placed onto the image using the position of the eyes and mid-upper lip vermilion border, with mask position refinement performed manually. (**B**) Lateral left side 2D photo prepared for wrinkle measurement at different regions. New site images were delineated via positioning of points at the lateral canthus of the left eye and the left corner of the mouth; the distance from the eye to the mouth was used to ensure correct sizing and positioning of each box. The upper lip was further segmented from the surrounding features in the box region surrounding the mouth using a point at the mid-upper lip vermilion border. (**C**) A masked image prepared for pigmented spot digital measurement, the line across the image represents where the image was additionally cropped for telangiectasia measurement on the cheeks and nose.

Telangiectasia

The 2D front photographs that had been previously masked for the pigmented spots measurement were used to measure red veins on the nose and cheeks. The images were further cropped down the face, removing the forehead (Figure 1C), using Adobe Photoshop CS4 (www.adobe.com). Differently to pigmented spots, telangiectasia almost solely present on the nose and cheeks.

Image analysis

All image analyses were conducted using MATLAB.

Wrinkles

First, large scale shading in the image was removed by flat-fielding the image – dividing the original image by a Gaussian filtered version of the image and then rescaling; the image was smoothed using Gaussian and median filters to remove fine skin texture and very small objects such as pores (Figure 2A-B). The 2nd derivative (which highlights dark ridges, Figure 2C) was used for a dual threshold technique inspired by the Canny Edge Detector algorithm. Low and high thresholds were applied separately using the red green channels for the high threshold and the red channel for the low threshold. Two new binary images containing candidate wrinkle areas were generated, with smaller finer wrinkles more commonly present in the low threshold image (Figure 2D-E). The candidate wrinkles in both images were accepted or rejected based on shape

(eccentricity and solidarity), intensity, and direction metrics. A line connection algorithm on the high threshold binary image was additionally performed (Figure 2E) to prevent rejection by the size of wrinkles broken into sections. Hence, line sections were connected if they were close to each other and pointing in a similar direction. The final detected wrinkles were taken from the low threshold binary image if they overlapped with part of a wrinkle in the high threshold image (Figure 2F). Wrinkles in the low threshold image were also included if they were not detected by the high threshold filtering but were very linear in nature (eccentricity threshold) and above a certain size.

Finally, a number of wrinkle variables were outputted: (1) Area, consisting of the cumulative number of pixels detected as wrinkles as a percentage of total skin area (i.e. the unmasked skin for full face wrinkles and the box area for localized site wrinkles). (2) Number, consisting of the total number of individual detected wrinkle lines, corrected for total skin area. (3) Length, consisting of the cumulative length of (skeletonized) areas detected as wrinkles, normalized by the square root of the total skin area. (4) Mean width, the average width of the detected wrinkles. (5) Depth, average of the 2nd derivative values for the pixels detected as wrinkles.

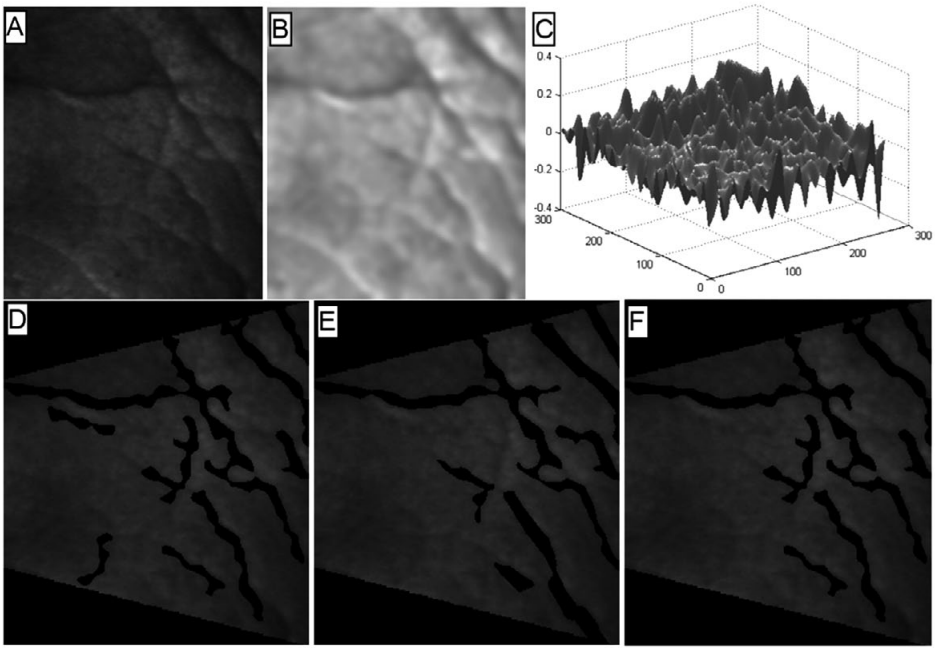

Figure 2. Illustration of dual threshold wrinkle detection on a crow's feet image. (**A**) Shows the original image; (**B**) is a flat-fielded and smoothed image; (**C**) a 3D representation of (B) which is an approximation of the 2nd derivative. The 2nd derivative detects bright and dark ridges in the image; dark ridges have positive values and correspond to wrinkles in the image. (**D**) Wrinkles detected by the low threshold (black lines), (**E**) wrinkles detected by the high threshold detection and (**F**) the final detected wrinkles – i.e. wrinkles in the low threshold image that intersect those in the high threshold image.

Pigmented spots

For the detection of pigmented spots and telangiectasia, we used the Difference of Gaussians technique on all three RGB channels. This algorithm uses a 2D Gaussian filter at two sizes to create new "contrast" images. A low-pass filter is used with a large standard deviation and a high-pass filter is used with a small standard deviation. The two filtered images from each RGB channel were subtracted and the resultant difference used to generate a contrast image (Figure 3B). Pigmented spots in the contrast image appear as blue spots (as the greatest contrast in their appearance to surrounding skin is in the blue channel). To further filter out spurious artifacts an intensity ratio threshold (targeting pixels with high blue values relative to their green and red values), a minimum pixel size (to remove noise), a solidarity threshold (to remove branched objects) and an eccentricity threshold (to remove linear objects – e.g. wrinkles) were applied to the contrast image, Figure 3B. The digital output of the detected blue spots consisted of two measures: (1) Area, consisting of the cumulative detected bluish and roundish areas, as a percentage of total skin area. (2) Number, consisting of the total number of individual detected areas, corrected for total skin area.

Telangiectasia

A contrast image was also created for detecting telangiectasia. Red/purple veins would appear green in color in the contrast weighted image, so an algorithm and threshold was applied to detect pixels with high green relative to red and blue values; additionally filtering was applied to target linear (eccentricity) and branched structures (solidarity), Figure 3C. As for pigmented spots, the digital output consisted of two measures: (1) Area, consisting of the cumulative detected greenish linear areas, as a percentage of total skin area. (2) Number, consisting of the total number of individual detected areas, corrected for total skin area.

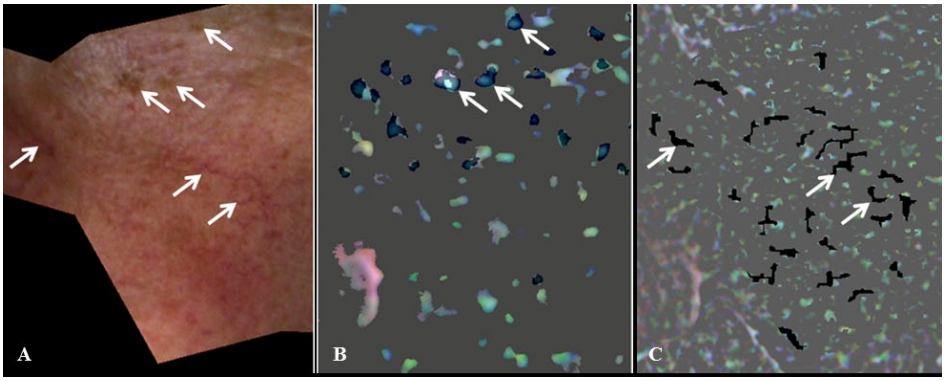

Figure 3. Illustration of pigmented spot and telangiectasia detection. (**A**) Shows the original image with pigmented spots (left facing arrows) and telangiectasia (right facing arrows); (**B**) is the contrasted image targeted to features approximate in size to pigmented spots, brown features appear blue. Detected spots are shaded; (**C**) is the contrasted weighted image targeted to features approximate in size to telangiectasia, red features appear as green and branched green objects were detected (black lines in image).

Statistical analysis

The intraclass correlation coefficient (ICC) was used to determine inter-rater reliability between the two investigators. A Two-Way Mixed model with the participants as a random factor and the raters as a fixed factor was used, with the ICC representing the reliability of the raters in the sample[28]. In case of a significant systematic difference in means between the two graders (i.e., someone graded consistently lower or higher), as tested by the paired-samples t-test, we used the absolute agreement type. Otherwise, we used the consistency type[28]. A correlation coefficient of ≥0.70 indicates a high reliability, 0.40–0.60 represents a moderate reliability and ≤0.3 a low reliability[29-31].

We calculated the Spearman's correlation coefficient (ρ) to describe the agreement between the average photonumeric grades (ordinal categorical variable) and the digital measurements (continuous variable). To interpret the similarity between the image analysis measures and photonumeric grading we used Colton's[32] recommendation of 0.25-0.50 to be fair, 0.50-0.75 to be moderate to good and >0.75 as very good to excellent. Men and women were analyzed separately as there appeared to be considerable differences between sexes. All analyses were performed using SPSS for Windows version 21.0 (SPSS, Chicago, IL, USA). A two-sided P-value of <0.05 was considered statistically significant.

RESULTS

Study population

All participants (N=150) were of northwestern European origin; the blinded comparisons between photonumeric and digital grading were based on a subgroup of 100 participants, with a mean age of 72.2 ±4.3 for the men and 71.4 ±3.7 for the women.

Photonumeric grading

The blinded inter-rater reliability of the photonumeric grading scales was good to excellent for all seven features. Full face wrinkles, pigmented spots and telangiectasia showed excellent ICCs (0.78-0.93). For wrinkle severity per site, the ICC was excellent for the forehead, crow's feet in men, nasolabial fold and upper lip (0.79-0.93), and good for crow's feet in women (0.65).

Digital measures

For the seven skin aging features, the mean affected area varied greatly, ranging from 0.6% for telangiectasia to 8.4% for crow's feet in men (Table 1). Detected wrinkles covered on average 5% of the face in both men and women, and covered more area on the forehead, crow's feet, and female upper lip (5.6% - 8.4%). However, upper lip wrinkles in men covered a notably smaller area (2.0%). Compared to the wrinkle features, the affected area of pigmented spots and telangiectasia was up to 10 times smaller. Although the photonumeric wrinkle grading for the localized facial

sites (i.e. forehead, crow's feet, nasolabial fold and upper lip) focused on the depth of the deepest wrinkle, the digital measure of depth (average depth of all wrinkles) did not give notably higher correlations than the digital area measure (e.g. Table 2).

Table 1. Means for the digital measures for all seven skin aging features and their correlations with average photonumeric grading, in men and women

	Men (N=50)		Women (N=50)	
Skin aging feature	**Mean ± SD**	**ρ**	**Mean ± SD**	**ρ**
Full face wrinkles	5.3 ± 2.2	0.79	5.2 ± 2.8	0.89
Forehead wrinkles	8.2 ± 6.5	0.63	6.9 ± 6.2	0.63
Crow's feet wrinkles	8.4 ± 5.0	0.52	5.6 ± 4.8	0.81
Nasolabial fold wrinkle	1.2 ± 1.0	0.86	0.6 ± 0.7	0.58
Upper lip wrinkles	2.0 ± 2.5	0.30	6.1 ± 6.4	0.76
Pigmented spots	0.8 ± 0.5	0.70	2.1 ± 1.0	0.69
Telangiectasia	0.6 ± 0.3	0.75	0.8 ± 0.5	0.60

Abbreviations: ρ, Spearman's correlation coefficient; SD, standard deviation.
Digital measures represent mean percentages of the affected area per total skin area. Spearman's correlation coefficients between the digital measures and photonumeric grading for each feature are given.

Table 2. Correlations between the different digital wrinkle measures outputted by the image analysis and average manual photonumeric grading for the crow's feet region

		Photonumeric grades	
Digital measures		**Depth**	**Number and depth**
Men	Number	0.40	0.62
	Depth	0.57	0.55
	Width	0.49	0.47
	Length	0.48	0.67
	Area	0.52	0.70
Women	Number	0.71	0.77
	Depth	0.58	0.59
	Width	0.55	0.53
	Length	0.80	0.86
	Area	0.81	0.86

Abbreviations: ρ, Spearman's correlation coefficient; Depth, average of 2 graders; Number and depth = (average number + average depth)/2.
The inclusions of wrinkle number as well as depth to the photonumeric scores increased the correlations, particularly for the digital number, length, and area measures.

Photonumeric grading vs. digital measures

Overall, the correlations between the photonumeric grading and digital measures were moderate to excellent for both sexes (ρ>0.50, P-value<0.001), except for upper lip wrinkles in men (ρ_m=0.30, P-value=0.035), Table 1. Full face wrinkle area gave excellent correlations with the photonumeric

grading in men and women (ρ_m=0.79 and ρ_w=0.89). The correlations between the photonumeric grading and the localized wrinkle area measures were excellent for nasolabial fold in the men, and upper lip and crow's feet in the women (ρ_m=0.86; ρ_w=0.76 and ρ_w=0.81, respectively), moderate-to-good for the forehead and crow's feet in the men, and for the forehead and nasolabial fold in the women (ρ_m=0.63 and 0.52, ρ_w=0.63 and 0.58, respectively), but fair for the upper lip in the men (ρ_m=0.30). A combined photonumeric score of wrinkle number and depth increased the correlations with the crow's feet digital area measure (ρ_m =0.52 to 0.70 and ρ_w=0.81 to 0.86, Table 2). For pigmented spots, there was a good correlation between photonumeric grading and the digital measures in both men and women – both approximately 0.7 (Table 1). The correlations for the digital telangiectasia area with the photonumeric grading were also good, particularly in the men (ρ_m=0.75 and ρ_w=0.60, Table 1).

The increase in the digital measures per increase in photonumeric grade was consistent for full face wrinkles, pigmented spots and telangiectasia (Figure 4). Per photonumeric grade, the digital measures significantly increased (Figure 4A) bar for pigmented spots grades 4-5 (Figure 4B) and telangiectasia grades 1-2 (Figure 4C).

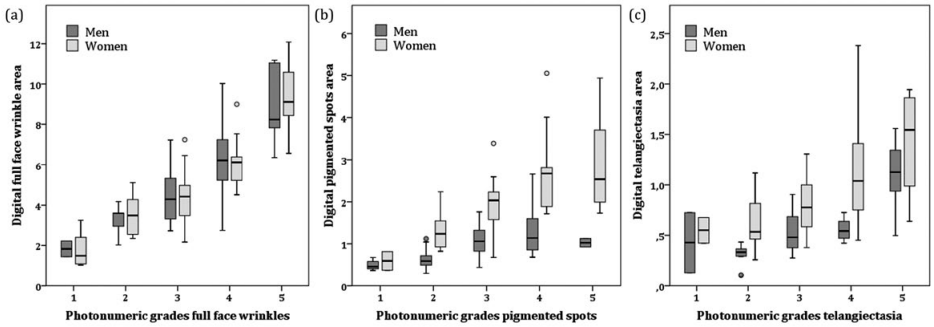

Figure 4. Boxplots of photonumeric vs. digital measures for three skin aging features, separately for men (N=50) and women (N=50). The average photonumeric grades (rounded up for half values) are shown on the x-axis, digital measures on the y-axis. The band in the box represents the median, with the bottom and top parts the first and third quartile. The bottom vertical line indicates data within 1.5 of the interquartile (IQR) range of the 1st quartile, and the top vertical line represents data within 1.5 IQR of the 3rd quartile. (**A**) Full face wrinkle measurement; (**B**) pigmented spots measurement; (**C**) telangiectasia measurement.

DISCUSSION

The digital area measures for wrinkles, pigmented spots and telangiectasia had moderate to excellent correlations with photonumeric grading, with the correlation for the upper lip wrinkle measure in men being the only exception.

Although there is no gold standard for photonumeric grading of the different components of skin aging, the good to excellent inter-rater reliability of our photonumeric scales suggests they are a valid comparative measurement for digital measures. The photonumeric full face wrinkle

scale, which was based on a combination of different wrinkle severity characteristics (i.e. number, length, width and depth), had higher correlations with the digital area measure than the photonumeric wrinkle grading from the localized wrinkle sites. This was likely due to the fact that the photonumeric grading for the localized sites graded the depth of the deepest wrinkle rather than overall wrinkle severity. A combined photonumeric score for crow's feet wrinkle number and depth gave higher correlations with the digital area measure than photonumeric depth alone, indicating that area was indeed a better measure of overall wrinkle severity than wrinkle depth. However, the digital depth measure did not have consistently higher correlations with photonumeric depth than the digital area did. This could be due to the fact that digital depth represented the average depth across all detected wrinkles rather than the depth of the deepest wrinkle. Hence, for future validation studies we recommend comparing the area of wrinkles detected with a photonumeric scale of overall wrinkle severity or, if depth of the deepest wrinkle is a research interest, adapting the image analysis techniques to generate a more similar digital measure.

All outcomes were stratified by sex because visible skin aging differs between men and women[18,24,25]. Although evaluating sex differences in skin aging warrants investigation in larger studies, we found sex differences in the correlations between the digital measures and photonumeric grading. The crow's feet and upper lip wrinkle measures in the men showed a much lower correlation than in the women. Male sex is an independent risk factor for sagging of upper eyelids[33], which can merge with crow's feet wrinkles. On inspection, eyelid sagging was found to be detected by the image analysis in some images, but was ignored by the graders. Hence, sagging eyelids in men could be reducing the correlation between digital wrinkle area and the photonumeric grading. As eyelid sagging and crow's feet wrinkles are likely two distinct phenotypes, distinguishing between the two features in future image analysis techniques will help isolate the risk factors specific to each.

The lowest correlation between the image analysis and photonumeric grading was for the upper lip in the men. On visual inspection of the images, we identified three main reasons. First, the men had very few wrinkles on the upper lip compared to the women; this sex difference has been confirmed in other studies[25]. This means that any error in the image analysis (e.g. missing the only wrinkle present) has a much larger impact on the digital measure. Second, the region of the upper lip used for digital measurement was small (see Figure 1B) compared to that used by the graders (full upper lip region) and the deepest wrinkle (which was the only one graded) lay outside the digital area for some male participants. Third, the presence of stubble in this region meant there were a few individuals where the darkness of the stubble facilitated the odd erroneous wrinkle detection. Hence, further optimization and validation of the upper lip wrinkle detection in men is required (e.g. to eliminate stubble effects and enlarge the lip area analyzed).

For the women, the nasolabial fold area correlation with the photonumeric grading was lower than for the men. On visual inspection of the detected nasolabial fold in the participant images, the women were found to have more surrounding wrinkles, which were occasionally detected by the image analysis as being part of the nasolabial fold; in such situations the human graders

would have excluded the presence of such wrinkles in their grading. To remove the influence of wrinkles in the region, images were filtered on position and angle of the nasolabial fold. This caused the lower percentage coverage of the nasolabial fold in this region compared to wrinkle coverage in other regions. However, refinement of the technique to further remove the influence of surrounding wrinkles would help improve this measure further, particularly for measurements in women.

Limitations to the study here include a lack of heterogeneity in the sample population, which was a middle-aged to elderly northwestern European sample. There was no corresponding increase in the digital measures between the highest two grades for pigmented spots or the lowest two grades for telangiectasia. Although the number of participants in the extreme grades was very low (<5), it suggests the digital measures might not be discriminating appropriately between these grades, which will be more common in older (for the pigmented spots) or younger (for telangiectasia) individuals. Hence, further image analysis optimization and validation are required before these techniques can be utilized with confidence in older or younger cohorts, and additionally for darker skinned individuals. The image analysis of wrinkles at the localized sites was only performed on the left side of the face. Hence, there may have been under- or overestimation of the amount of wrinkles due to asymmetry in facial photoaging[34,35]. However, at a population level it probably does not radically influence the results. Finally, although image analysis techniques are consistently applied to every image, technical variation in the images can bias the outcomes. The Premier 3dMD face3-plus UHD camera was designed for analysis of facial structure via 3D rendering rather than image analysis on the 2D camera images. Hence, there was no face rest resulting in skin luminance variability across participants. To counteract such effects, the image analysis methods incorporated compensatory algorithms such as utilizing the contrast in color and lightening (e.g. 2nd derivative) within the images rather than absolute color or lightening values. Thus, the digital measures should have been unaffected by differences in lightening levels, although they would still be affected by variations in color balance and the total contrast in light intensity. Hence, more standardized camera set-ups and greater image resolution should improve the reproducibility of the image analysis techniques in the future.

Although previously the measurement of skin aging has been mainly based on photonumeric scales[2,6,7,36], digital measurement has enough advantages over photonumeric grading to suggest it will become the main choice in the future. First, there were good to excellent correlations for the majority of digital measures with photonumeric grading. Second, digital measurement generates a continuous outcome giving more statistical power to detect risk factor associations[37]. Third, better quality images, more automated masking, improved lightening consistency etc. will further improve the utility of image analysis techniques in the future. Finally, digital measurement is less time consuming once an image analysis system is built as it can calculate multiple outcomes per aging component and measure multiple features almost simultaneously.

In conclusion, our digital grading system has proven to be a suitable scale for the measurement of wrinkles (with upper lip wrinkles in men being the exception), pigmented spots and telangi-

ectasia. Digital measurement provides continuous outcomes for different aspects of skin aging, which makes it useful for unbiased discrimination of feature differences in photographic images. Thus, these digital measurement systems for skin aging features demonstrate potential for use in observational and experimental skin aging research.

ACKNOWLEDGMENTS

The authors are grateful to the study participants, the staff from the Rotterdam Study and the participating general practitioners and pharmacists. We thank Sophie Flohil, Emmilia Dowlatshahi, Robert van der Leest, Joris Verkouteren, Ella van der Voort and Shmaila Talib for collecting the phenotypes. Additionally we thank Sophie van den Berg for masking and reviewing all the photographs. We would like to acknowledge Peter Murray for advice around statistical analyses and Arthur Weightman for building the software to segment the localized facial sites.

REFERENCES

1. The American Society for Aesthetic Plastic Surgery Reports Americans Spent Largest Amount on Cosmetic Surgery Since The Great Recession of 2008. Statistics, Surveys & Trends2014.
2. Griffiths CE, Wang TS, Hamilton TA, Voorhees JJ, Ellis CN. A photonumeric scale for the assessment of cutaneous photodamage. *Arch Dermatol*. 1992;128(3):347-51.
3. Larnier C, Ortonne JP, Venot A, Faivre B, Beani JC, Thomas P, et al. Evaluation of cutaneous photodamage using a photographic scale. *Br J Dermatol*. 1994;130(2):167-73.
4. Lemperle G, Holmes RE, Cohen SR, Lemperle SM. A classification of facial wrinkles. *Plast Reconstr Surg*. 2001;108(6):1735-50; discussion 51-2.
5. Rzany B, Carruthers A, Carruthers J, Flynn TC, Geister TL, Gortelmeyer R, et al. Validated composite assessment scales for the global face. *Dermatol Surg*. 2012;38(2 Spec No.):294-308.
6. Carruthers A, Carruthers J. A validated facial grading scale: the future of facial ageing measurement tools? *J Cosmet Laser Ther*. 2010;12(5):235-41.
7. Guinot C, Malvy DJ, Ambroisine L, Latreille J, Mauger E, Tenenhaus M, et al. Relative contribution of intrinsic vs extrinsic factors to skin aging as determined by a validated skin age score. *Arch Dermatol*. 2002;138(11):1454-60.
8. Vierkotter A, Ranft U, Kramer U, Sugiri D, Reimann V, Krutmann J. The SCINEXA: a novel, validated score to simultaneously assess and differentiate between intrinsic and extrinsic skin ageing. *J Dermatol Sci*. 2009;53(3):207-11.
9. Monestier S, Gaudy C, Gouvernet J, Richard MA, Grob JJ. Multiple senile lentigos of the face, a skin ageing pattern resulting from a life excess of intermittent sun exposure in dark-skinned caucasians: a case-control study. *Br J Dermatol*. 2006;154(3):438-44.
10. Morizot F. LS, Guinot C. et al. Development of photographic scales documenting features of skin ageing based on digital images. *Ann Dermatol Venereol*. 2002;129 (hors-serie 1, cahier 2):1S402.
11. Suppa M, Elliott F, Mikeljevic JS, Mukasa Y, Chan M, Leake S, et al. The determinants of periorbital skin ageing in participants of a melanoma case-control study in the U.K. *Br J Dermatol*. 2011;165(5):1011-21.
12. Chung JH, Lee SH, Youn CS, Park BJ, Kim KH, Park KC, et al. Cutaneous photodamage in Koreans: influence of sex, sun exposure, smoking, and skin color. *Arch Dermatol*. 2001;137(8):1043-51.
13. Tanghetti EA. Split-face randomized treatment of facial telangiectasia comparing pulsed dye laser and an intense pulsed light handpiece. *Lasers Surg Med*. 2012;44(2):97-102.
14. Kennedy C, Bastiaens MT, Bajdik CD, Willemze R, Westendorp RG, Bouwes Bavinck JN, et al. Effect of smoking and sun on the aging skin. *J Invest Dermatol*. 2003;120(4):548-54.
15. Hatzis J. The wrinkle and its measurement--a skin surface Profilometric method. *Micron*. 2004;35(3):201-19.
16. Jacobi U, Chen M, Frankowski G, Sinkgraven R, Hund M, Rzany B, et al. In vivo determination of skin surface topography using an optical 3D device. *Skin Res Technol*. 2004;10(4):207-14.
17. Jiang LI, Stephens TJ, Goodman R. SWIRL, a clinically validated, objective, and quantitative method for facial wrinkle assessment. *Skin Res Technol*. 2013;19(4):492-8.
18. Luebberding S, Krueger N, Kerscher M. Quantification of age-related facial wrinkles in men and women using a three-dimensional fringe projection method and validated assessment scales. *Dermatol Surg*. 2014;40(1):22-32.
19. Luebberding S, Krueger N, Kerscher M. Comparison of Validated Assessment Scales and 3D digital fringe projection method to assess lifetime development of wrinkles in men. *Skin Res Technol*. 2014;20(1):30-6.

20. Gossage KW, Weissman J, Velthuizen R. Segmentation of hyper-pigmented spots in human skin using automated cluster analysis. *Proc SPIE*. 2009;7161.
21. Miyamoto K, Takiwaki H, Hillebrand GG, Arase S. Development of a digital imaging system for objective measurement of hyperpigmented spots on the face. *Skin Res Technol*. 2002;8(4):227-35.
22. Stamatas GN, Balas CJ, Kollias N. Hyperspectral image acquisition and analysis of skin. *Proc SPIE*. 2003;4959.
23. Hofman A, Darwish Murad S, van Duijn CM, Franco OH, Goedegebure A, Ikram MA, et al. The Rotterdam Study: 2014 objectives and design update. *Eur J Epidemiol*. 2013;28(11):889-926.
24. Tsukahara K, Hotta M, Osanai O, Kawada H, Kitahara T, Takema Y. Gender-dependent differences in degree of facial wrinkles. *Skin Res Technol*. 2013;19(1):e65-71.
25. Paes EC, Teepen HJ, Koop WA, Kon M. Perioral wrinkles: histologic differences between men and women. *Aesthet Surg J*. 2009;29(6):467-72.
26. Roland Bazin ED. Skin Aging Atlas. Paris: Editions MED'COM; 2007. 103 p.
27. Daniell HW. Smoker's wrinkles. A study in the epidemiology of "crow's feet". *Ann Intern Med*. 1971;75(6):873-80.
28. McGraw KO WS. Forming Inferences About Some Intraclass Correlation Coefficients. *Psychological Methods*. 1996;1(1):30-46.
29. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420-8.
30. Nunnally JC BI. Psychometric theory. New York: McGraw-Hill Inc.; 1994.
31. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol*. 2007;60(1):34-42.
32. Colton T. Statistics in Medicine. Boston, MA: Little, Brown and Company; 1974. p 211 p.
33. Jacobs LC, Liu F, Bleyen I, Gunn DA, Hofman A, Klaver CC, et al. Intrinsic and extrinsic risk factors for sagging eyelids. *JAMA Dermatol*. 2014;150(8):836-43.
34. Mac-Mary S, Sainthillier JM, Jeudy A, Sladen C, Williams C, Bell M, et al. Assessment of cumulative exposure to UVA through the study of asymmetrical facial skin aging. *Clin Interv Aging*. 2010;5:277-84.
35. Pierard GE, Hermanns-Le T, Gaspard U, Pierard-Franchimont C. Asymmetric facial skin viscoelasticity during climacteric aging. *Clin Cosmet Investig Dermatol*. 2014;7:111-8.
36. Honeck P, Weiss C, Sterry W, Rzany B, Gladys study g. Reproducibility of a four-point clinical severity score for glabellar frown lines. *Br J Dermatol*. 2003;149(2):306-10.
37. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med*. 2006;25(1):127-41.