Published in final edited form as:
Arch Dermatol. 2011 January ; 147(1): 31–36. doi:10.1001/archdermatol.2010.392.

Development of a photographic scale for consistency and guidance in dermatological assessment of forearm sun damage

NE McKenzie, **K Saboda**, **LD Duckett**, **R Goldman**, **C Hu**, and **CN Curiel-Lewandrowski**
Arizona Cancer Center, University of Arizona, Tucson, 85724-5024, USA.

Abstract

Objectives—To develop a photographic sun damage assessment scale in forearm skin and assess its feasibility in assuring consistent identification, description and quantification of sun damage when used as part of an ongoing Dermatologic Assessment Scoring system in clinical studies.

Design—Blinded comparison. Ninetysix standardized 8×10 digital photographs of participant forearms were taken. The photographs were then graded by our expert dermatologist using the existing clinical sun damage assessment scale used in clinical trials until all categories contained photographs representative of each clinical sign. The expert dermatologist's set of grades created the criterion standard. The same photos were put into binders in triplicate in random arrangement and were provided to five blinded community dermatologists who rated them using the clinical sun damage assessment scale.

We took standardized 8×10 digital photographs of participant forearms and graded them into scoring categories by clinical sign using our existing clinical sun damage assessment scale until all categories were saturated for each clinical sign. Initial selection and grading were performed by our designated dermatologist expert to create a criterion standard. Three binders each containing 96 randomly arranged photos were graded by 5 blinded community dermatologists and a second reading by our criterion standard dermatologist.

Setting—Academic skin cancer prevention clinical trials clinic with a high level of experience assessing sundamaged skin.

Participants—Convenience sample consisting of any adult over 18 years of age including participants taking part in screenings, chemoprevention, and/or biomarker studies. Six community and academic dermatologists were recruited to grade the photographs.

Main Outcome Measure(s)—Reproducibility and agreement of grading among dermatologists as assessed by Spearman's correlation coefficient to assess the correlation of scores given for the same photograph, kappa statistics for ordinal data, and variability of scoring among dermatologists using ANOVA models with evaluating physician and photos as main effects and interaction effect variables to account for the difference in scoring among dermatologists.

Results—The correlations (~70% to above 90%) between dermatologists were all statistically significant ($p < 0.0001$) Scores showed good to substantial agreement but were significantly

Corresponding author: Dr. Naja McKenzie nmckenzie@azcc.arizona.edu.

different (p<0.0001) for each of the four clinical signs and the difference varied significantly (p<0.0001) from photo to photo.

Conclusions—The use of the photographic sun damage assessment scale is highly feasible with an acceptable inter-observer agreement. Our findings also suggest that training will be useful in improving physician agreement on grading of the photographs.

Introduction

The overall goal of our program project grant is to develop safe, effective, and cost-efficient strategies for the secondary prevention of skin cancer. Primary prevention strategies have failed adequately to reduce skin cancer incidence and morbidity, especially in the Southwestern United States. New strategies are therefore necessary to reduce the public health burden of this disease. Our chemoprevention and biomarker protocols rely on consistent clinical identification, description and quantification of sun damage in forearm skin. To date, no valid and reliable photographic assessment scale of forearm skin sun damage has been developed. It was the purpose of this study to develop a photographic assessment scale of sun damage in forearm skin that can be used to assure consistent identification, description and quantification of sun damage when documented as part of our ongoing Dermatologic Assessment Scoring system in chemoprevention and biomarkers studies.

Objectives

The clinical assessment of human skin for sun damage is an essential but highly subjective process for evaluating skin cancer risk. Clinical assessment is also a vital part of evaluating the effectiveness of agents being tested for their potential ability to reverse sun damage. Given that our chemoprevention trials utilize a histopathological scoring system along with clinical evaluation to assess efficacy of test articles, biopsies are required for evaluation. Human subject considerations therefore suggest that forearm skin, rather than facial skin, should be used to test the safety and efficacy (phase 1 and 2 trials) of chemopreventive agents for the reversal of sun damage. This consideration alone makes an objective grading scale for the measurement of sun damage in forearm skin essential to a successful chemopreventive agent development program. Furthermore, a standardized teaching set will be valuable for developing a reproducible method and can support the comparison of findings from a variety of studies in our skin cancer chemoprevention program. To date, no standardized photographic scale exists for assessment of forearm sun damage. The goal of this study was to develop such a scale for use in research (or practice) where assessment of photodamage is integral.

Methods

Forearm photodamage assessment in our prior and current studies is performed using a subjective 10-point scale for each of five clinical signs of UV induced skin damage: fine wrinkling, coarse wrinkling, mottling, hyperpigmentation, and a global assessment. The global assessment is used by dermatologists to give their overall impression of sun damage. Each clinical sign is ranked and subdivided as follows: Absent (0), Mild (1-3), Moderate

(4-6), and Severe (7-9). This approach is similar to the R.W. Johnson Pharmaceutical Research Institute scale[1, 2] which is used for assessment of photodamage in facial skin, but our scale omits any elements that specifically refer to facial photodamage. The scale is presented in Table 1.

Participants

In the spring and summer of 2007, forty-eight adults over 18 years of age [female = 26 (54.2%) averaging 52 years of age, male = 22 (45.8%) averaging 63 years of age]. Participants identified themselves as Caucasian (n=47) and African American (n=1). Ethnically, participants identified as Hispanic (n=6) and non-Hispanic (n=36) while 6 did not provide any ethnic identification. The sample consented for this project included community volunteers and participants taking part in screenings, chemoprevention or biomarkers studies at the Skin Cancer Prevention Annex,. Individuals whose dorsal forearms were unsuitable for use in a photographic scale, including those with significant inflammation or irritation, tattoos or other markings, were not eligible. Some candidate subjects were invited to participate on the basis of having significant sun damage previously evaluated in our clinic. Others responded to word-of-mouth recruitment among friends, family and associates.

A criterion standard is described by the Journal of the American Medical Association (JAMA) as the term the journal prefers over “gold standard.” It signifies a standard for comparison of a new screening test or diagnostic evaluation The term can also be used to indicate a performance standard to which experts or peers agree and to which individual practice can be compared [3].

One academic and five community dermatologists agreed to assist as raters. Of these, the academic physician was designated as the project's expert dermatologist. This dermatologist is the primary study physician leading our clinical trials and therefore has the most experience assessing forearm skin photodamage. This physician’ initial grading was designated as the criterion standard for subsequent gradings using the photographic scale.

Digital photography

Digital photographs were taken of the dorsal forearms from knuckle (metacarpalphalangeal) to elbow. The right and left forearms were photographed for a total of 96 unique photos. A Nikon Coolpix 4300 digital camera was used with standardized lighting, background and positioning methodology to insure consistency. The expert dermatologist graded the photographs into score categories by clinical sign using our existing clinical sun damage assessment scale until all score categories were saturated for each clinical sign. Individuals on the extremes, almost no sun damage and very severe sun damage, had to be sought out by referral.

Digital photographs were stored with a unique numeric identifier. To avoid personal identification due to jewelry or identifying marks, photos were cropped to include knuckle to elbow only. No other alterations or enhancements were made to the photographs.

Randomization and grading

Each photograph was printed unedited on photo paper, coded with its unique identifier, and paired with a blank dermatological assessment scale form (Table 1). The expert dermatologist performed the initial grading of the photos thus establishing our criterion standard for comparison as shown in Table 2.

Each photo page was then reprinted to create triplicate image sets, resulting in full image evaluation sets of 288 photo pages. The pages were randomly ordered using Stata-generated random numbers and placed into binders. Identical image sets were delivered to 5 evaluating dermatologists who each blindly evaluated the 96 unique photos three times according to the dermatologic assessment scale shown in Table 1. Finally, the dermatologist designated as the criterion standard performed a repeat evaluation of the randomized set of photos.

Each dermatologist should have reviewed 288 photographs. After the binders were returned it was discovered that two dermatologists received an incomplete folder with one photograph missing. Thus each of these two dermatologists was only able to evaluate 287 photographs. The two missed evaluations are treated as missing data. For ANOVA analyses to be conducted a missing value was imputed using average of available data for the same reviewer and photograph.

Analysis and Results

The non-parametric Spearman ρ correlation coefficient was used to study the correlation of all scores given for the each photograph. ANOVA models with random effects were employed to study the difference in assessments by different dermatologists. All analyses and graphs were carried out in Stata version 10 (StataCorp LP, College Station, Texas).

Criterion standard

We first investigated the relationship among the four sun damage scores given to each photograph by the expert dermatologist (one as the standard and three as assessments of the binders three months later). Table 3 summarizes the non-parametric Spearman's rho correlation coefficients. All correlations were near or above 90% and all p-values <0.0001 Thus the expert dermatologist's assessments of the same photos over time were highly correlated near or above 90% for all four clinical signs and all p-values were <0.0001.

Assessment by Community Dermatologists

The Spearman's correlation between the expert dermatologist's assessment and the scores given by the five community dermatologists ranged from around 70% to above 90%. They were all statistically significant with p values < 0.0001. The Spearman correlations between expert dermatologist assessment and the assessments by the five community dermatologists are given in Table 4. The correlations ranged from a low of 73% to above 90% and all are statistically significant with p values < 0.0001. These results show that assessments by all dermatologists had a strong linear relationship with the criterion standard scores.

However, strongly correlated scores can be quite different in magnitude and ultimately fail to show agreement. Therefore, in order to quantify agreement among the community dermatologists and the criterion standard, we calculated the kappa statistic for ordinal data (STATA) [4]. Calculation of kappa is based on the ratio of the observed agreement to the expected (i.e. by chance) agreement. Kappa statistics, shown in Table 6, all fell between 0.28 and 0.76. Guidelines for interpretation of kappa vary. Landis and Koch [5] would categorize the former as "fair" and the latter as "substantial".

Agreement among raters and the criterion standard can also be expressed in terms of percent agreement which is calculated as part of the kappa statistic. Expressed in this manner, raters agreed with the criterion standard from 71 to 92% of the time. The highest percent agreement was between the original and final, blinded rating session of the expert dermatologist.

Figures 1a-1d show the distribution of maximum deviation from the criterion standard for each dermatologist and each clinical sign. Deviation is defined as the difference between a given score and the criterion standard, and the maximum deviation is the one with the greatest magnitude (positive or negative) among the three scores for each photograph. We see that a deviation of ±3 was not rare, and sometimes the magnitude of deviation could exceed 5.

We used two-way ANOVA analysis to examine the dermatologist effect and the photo effect. Both of these are main effects and their interactions are considered random. The two missing data points were imputed using the average of the other two scores by the same dermatologist for the same photo. To calculate ANOVA,aAll of the expert dermatologist's assessments were excluded from the data to avoid any potential bias. ANOVA indicated that the scores given by the five remaining dermatologists were significantly different ($p<0.0001$) for each of the four clinical signs, and the difference tended to vary from photo to photo ($p<0.0001$).

Discussion

Our clinical trials focus primarily on chemoprevention of non-melanoma skin cancer with topical drugs being applied to the forearm. To a significant extent, current clinical protocols rely on consistent clinical identification, description and quantification of sun damage in forearm skin to evaluate baselines and efficacy. To date, no valid and reliable photographic assessment scale of forearm skin sun damage has been developed. The purpose of this study was to develop and test a photographic assessment scale that can be used to assure such consistency when documented by study dermatologists as part of our ongoing Dermatologic Assessment Scoring system in clinical studies. We anticipated that the scale would be used as a reference and training set in order to document and build consistency in our clinicians' evaluation of forearm sun damage.

The quest for consistency in clinical assessment of sun damage has led to the development of various objective grading methods which allow for characterization as well as quantification of sun damage. The methods include descriptive grading scales [1, 6], visual analogue scales [7, 8] and photographic grading scales [6, 9] . Weiss and colleagues developed a descriptive scale for the assessment of overall cutaneous photo-aging to be used along with photographic samples [6] , but did not discuss measures pertaining to agreement and validity. The R.W. Johnson descriptive scale [1] includes a detailed description of the manifestations of sun damage with a chance-corrected agreement (κ coefficient) of 0.11. Chance–corrected agreement ranges from −1 to +1 with scores of .4 to .75 considered fair and >.75 excellent or substantial [5, 10]. This scale is similar to our own Clinical Assessment Scale, except that it is intended for facial skin. Furthermore, on our scale, hyperpigmentation and mottling have been combined to a single clinical sign and renamed Abnormal Pigmentation as, in the opinion of all of our principal investigators, there is not sufficient difference between hyperpigmentation and mottling in our target population to justify two separate clinical signs.

Visual analog scales rely on clinicians to estimate features visually on a metrically defined horizontal line. Developers of visual analog scales for assessment of sun damage [7, 8] have described these instruments as more sensitive than descriptive scales and highly reproducible, but have not reported chance-corrected agreement or repeatability. Our 10-step Clinical Assessment Scale consists of 3 levels of severity, mild, moderate and severe. Each of these is then subdivided into 3 numerical grades, thus allowing for a more nuanced scale, not unlike a visual scale.

Photographic scales have the advantage of providing a consistent visual frame of reference thus minimizing variability in perception and subjectivity. Common points of reference increase consensus on dermal assessment and allow development of a common understanding of the extent of damage signified by each of the clinical signs on the scale. Larnier's 6-point photographic scale [9] consisted of a set of 3 standardized photos to represent each of 6 grades of sun damage ranging from mild to very severe. The photos were taken in a standard manner, from the same angle and of the same side (left) and region of the face. On assessment of inter-observer agreement, chance-corrected κ scores ranged from 0.44 to 0.76 on first and second occasions. In addition, dermatologists with and without experience with sun damaged skin scored similarly, supporting the notion that a photographic scale increases objectivity and standardization. An upper extremity photonumeric scale was developed to assess skin aging in smokers and non-smokers on the upper inner arm considered protected from the sun [11]. The scale was effective in showing greater skin aging in smokers than non-smokers .

Our findings support the ability of blinded dermatologists without prior training in use of the scale and not clinically associated, to achieve good to excellent agreement and strong linear correlation among their scores as well as internal consistency of ratings, all at a level of high statistical significance. Nevertheless, there were differences in how the dermatologists rated the photographs. All dermatologists have similar years of experience and we cannot immediately explain the differences in how the community dermatologists rated the photos, although one sees primarily a retiree population and did rate the photos less severely. The size of maximum differences may be related to the type of patients normally seen in the practices of the community dermatologists. However, even without training our dermatologists achieved high agreement and significant correlation in how they rated the photos. The high percent agreement testifies to the potential for improvement in consistency with training among dermatologists for whom agreement is vital.

Conclusions

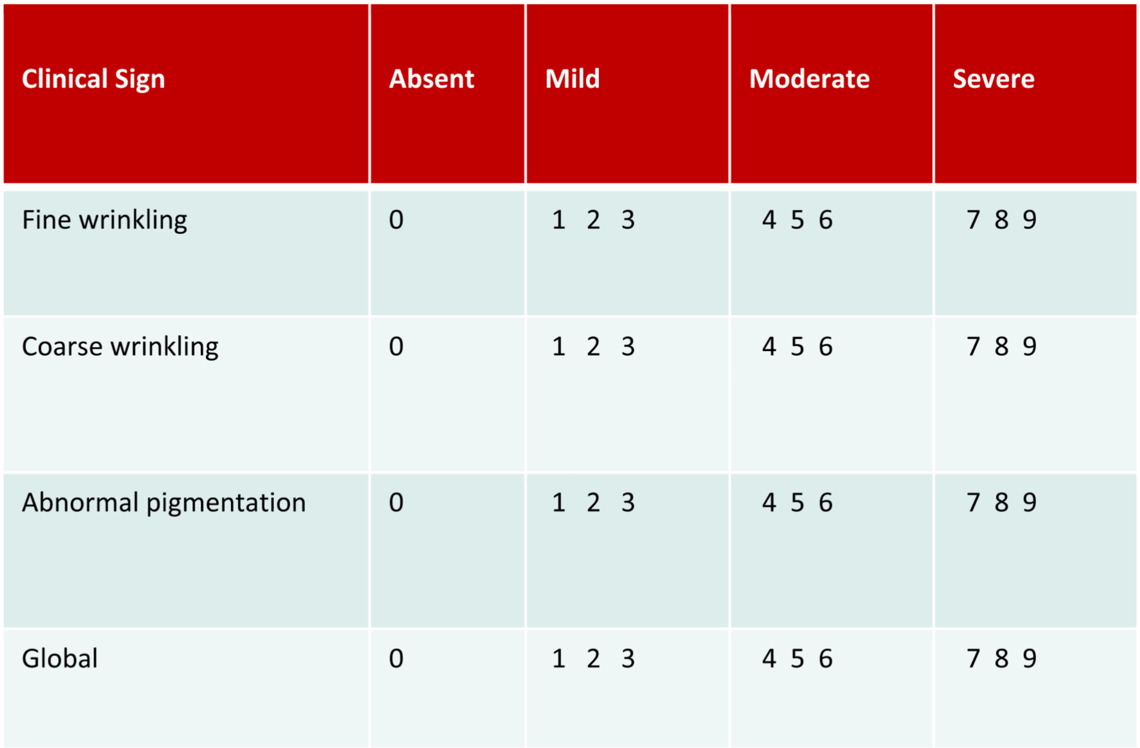
Based on these results, the expanded Dermatologic Assessment Form and Photographic Scale have great potential to yield highly consistent scoring of forearm sun damage in study participants. Further steps are needed to create a traiing set. Our expert dermatologist will select a few photographs that best represent each severity grade for each clinical sign. That image set will then be tested in our new program dermatologists to evaluate a baseline for agreement. We will then commence training of the group and repeat testing to evaluate improvement in agreement and correlation. We anticipate making the scale available for general clinical use in the future.

Acknowledgments

Grant funding: This study was supported in part by grant number, P01CA027502 from the National Institutes of Health.

References

1. Griffiths CE, Wang TS, Hamilton TA, Voorhees JJ, Ellis CN. A photonumeric scale for the assessment of cutaneous photodamage.[see comment]. Arch Dermatol. 1992; 128(3):347–351. [PubMed: 1550366]
2. Olsen EA, Katz HI, Levine N, et al. Sustained improvement in photodamaged skin with reduced tretinoin emollient cream treatment regimen: Effect of once-weekly and three-times-weekly applications. J Am Acad Dermatol. 1997; 37(2, Part 1):227–230. [PubMed: 9270508]
3. Journal of the American Medical Association. Author Instructions.
4. Jakobsson U, Westergren A. Statistical methods for assessing agreement for ordinal data. Scand J Caring Sci. Dec; 2005 19(4):427–431. [PubMed: 16324069]
5. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33(1):159–174. [PubMed: 843571]
6. Weiss JS, Ellis CN, Goldfarb MT, Voorhees JJ. Tretinoin therapy: practical aspects of evaluation and treatment. J Int Med Res. 1990; 18(3)
7. Lever L, Kumar P, Marks R. Topical retinoic acid for treatment of solar damage. Br J Dermatol. 1990; 122(1):91–98. [PubMed: 2404514]
8. Marks R, Edwards C. The measurement of photodamage. Br J Dermatol. 1992; 41:7–13. [PubMed: 1390188]
9. Larnier C, Ortonne JP, Venot A, et al. Evaluation of cutaneous photodamage using a photographic scale. Br J Dermatol. 1994; 130(2):167–173. [PubMed: 8123569]
10. Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics. 1977; 33(2):363–374. [PubMed: 884196]
11. Helfrich YR, Yu L, Ofori A, et al. Effect of smoking on aging of photoprotected skin: evidence gathered using a new photonumeric scale. Archives of Dermatology. Mar; 2007 143(3):397–402. [PubMed: 17372106] [erratum appears in Arch Dermatol. 2007 May;143(5):633].



Clinical Sign	Absent	Mild	Moderate	Severe
Fine wrinkling	0	1 2 3	4 5 6	7 8 9
Coarse wrinkling	0	1 2 3	4 5 6	7 8 9
Abnormal pigmentation	0	1 2 3	4 5 6	7 8 9
Global	0	1 2 3	4 5 6	7 8 9

Figure 1.
Dermatologic Assessment Form Forearm Photographic Assessment Scale

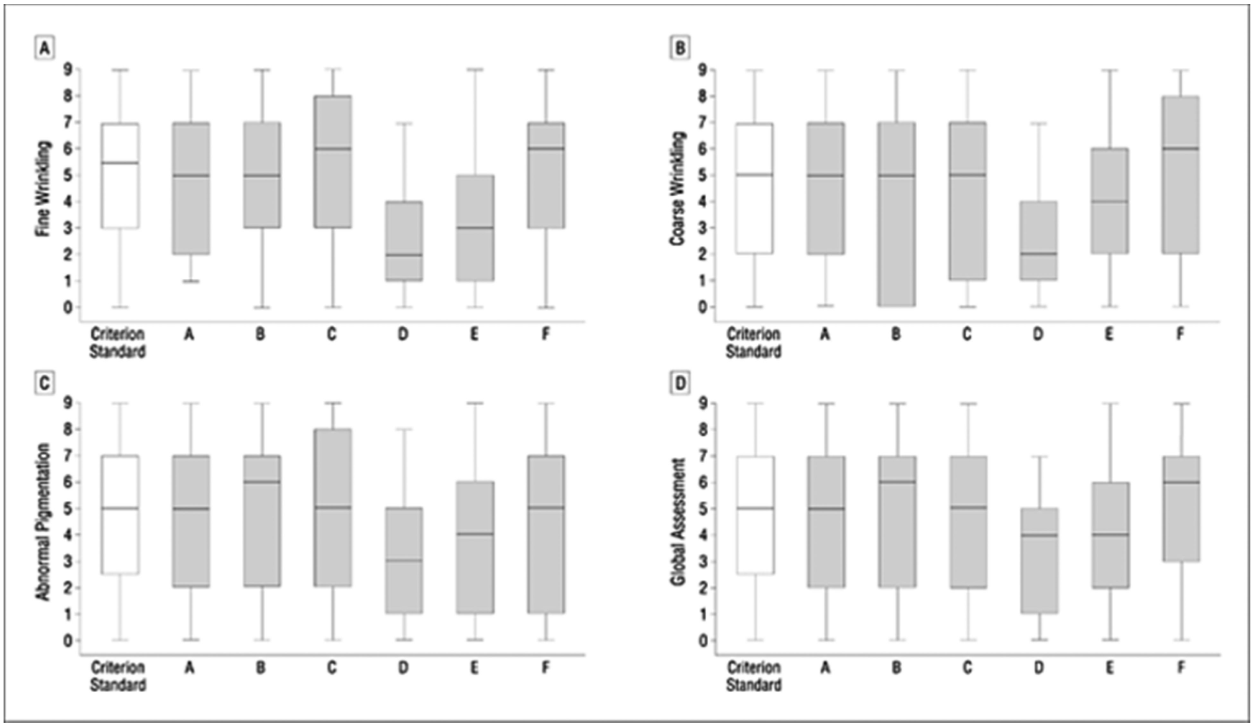


Figure 2.

A, Distribution of maximum fine wrinkling scoring deviation from the reference standard for each dermatologist (A is the reference standard dermatologist; B-F are the community dermatologists). B, Distribution of maximum coarse wrinkling scoring deviation from the reference standard for each dermatologist. C, Distribution of maximum abnormal pigmentation scoring deviation from the reference standard for each dermatologist. D, Distribution of maximum global assessment scoring deviation from the reference standard for each dermatologist.

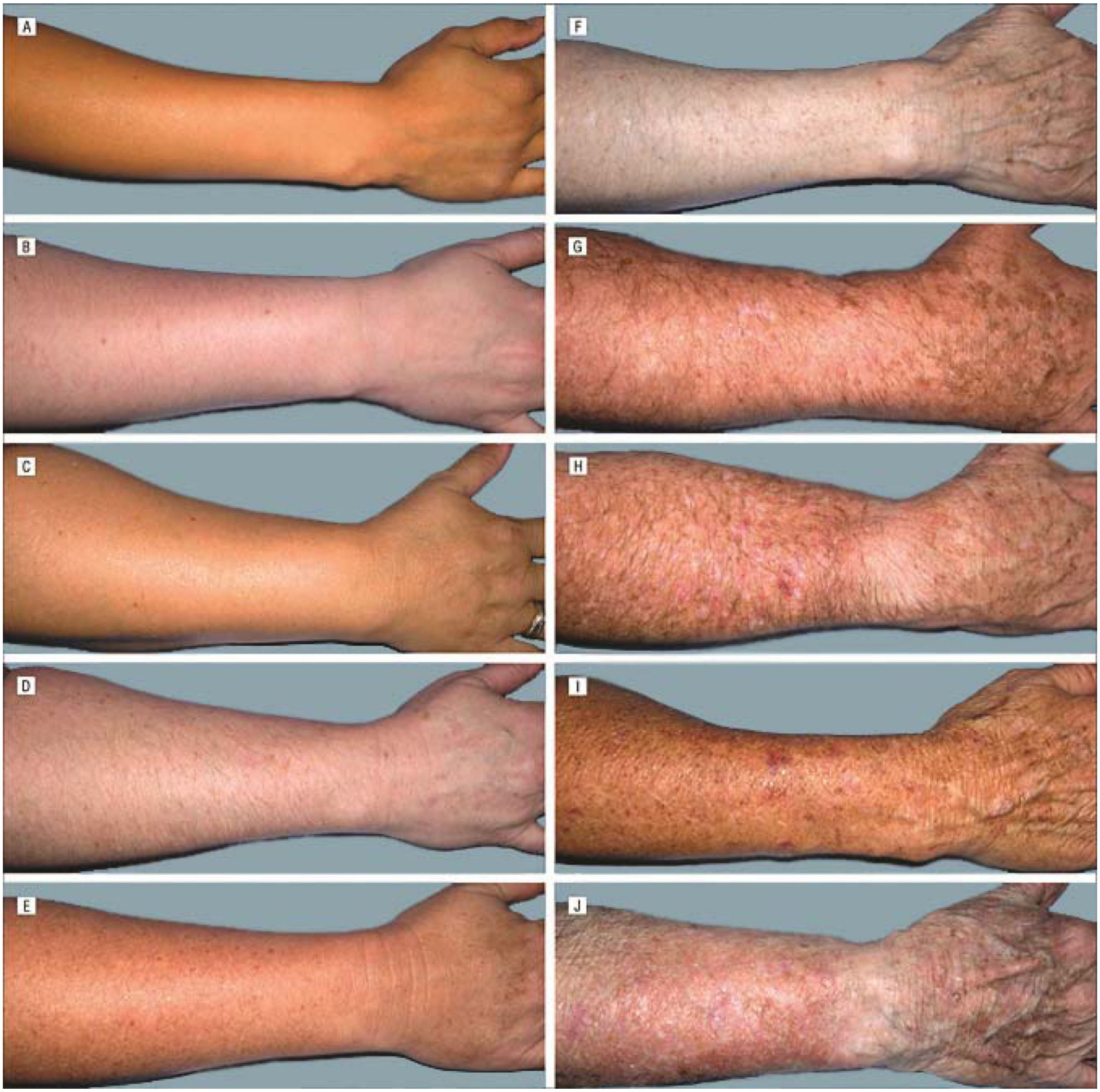


Figure 3.

Global assessment: severity score 0 (A); 1 (B); 2 (C); 3 (D); 4 (E); 5 (F); 6 (G); 7 (H); 8 (I); and 9 (J).

Table 1

Dermatologic Assessment Scale of Forearm Sun Damage

CLINICAL SIGN	ABSENT	MILD	MODERATE	SEVERE
Fine Wrinkling	0	1 2 3	4 5 6	7 8 9
Coarse Wrinkling	0	1 2 3	4 5 6	7 8 9
Abnormal pigmentation	0	1 2 3	4 5 6	7 8 9
Global	0	1 2 3	4 5 6	7 8 9

Table 2

Distribution of Criterion Standard Initial Grading by Category and Clinical Sign

Category	Level	Fine Wrinkling n	Coarse Wrinkling n	Abnormal Pigmentation n	Global Assessment n
None	**0**	2	9	5	6
Low	**1**	12	5	12	10
	2	7	11	7	8
	3	11	8	10	9
Moderate	**4**	6	5	4	5
	5	10	14	13	14
	6	21	10	19	17
Severe	**7**	15	19	13	15
	8	9	7	10	6
	9	3	8	3	6

Table 3

Correlation of Criterion Standard to repeated screening by expert dermatologist at 3 months*

Dermatologic Assessment Clinical Sign	Spearman ρ Correlation Coefficients		
	Image Set 1	Image Set 2	Image Set 3
Fine Wrinkling	0.87	0.92	0.91
Coarse Wrinkling	0.92	0.91	0.91
Abnormal Pigmentation	0.91	0.90	0.91
Global Assessment	0.92	0.93	0.93

All correlations are statistically significant at p <0.0001

Table 4

Correlation with Criterion Standard to Community Dermatologists[*]

Dermatologic Assessment Form Criteria		Spearman Correlation Coefficients		
	DERMATOLOGIST	**SET 1**	**SET 2**	**SET 3**
Fine Wrinkling	**EE**	0.79	0.87	0.87
	GG	0.88	0.89	0.88
	LI	0.71	0.69	0.74
	SS	0.81	0.81	0.83
	RM	0.74	0.86	0.86
Coarse Wrinkling	**EE**	0.90	0.89	0.91
	GG	0.92	0.93	0.91
	LI	0.82	0.83	0.85
	SS	0.82	0.83	0.87
	RM	0.86	0.85	0.88
Abnormal Pigmentation	**EE**	0.88	0.86	0.92
	GG	0.91	0.91	0.89
	LI	0.89	0.89	0.89
	SS	0.86	0.92	0.90
	RM	0.89	0.85	0.91
Global Assessment	**EE**	0.90	0.90	0.93
	GG	0.92	0.92	0.92
	LI	0.90	0.90	0.91
	SS	0.86	0.85	0.89
	RM	0.90	0.88	0.92

[*] **All correlations are statistically significant at p <0.0001**

Table 5

Kappa Statistics by Clinical Sign for Average Rater specific agreement vs. Criterion Standard

Fine Wrinkling	**Agreement**	**Kappa**	**95% confidence Interval for Kappa**		**Coarse Wrinkling**	**Agreement**	**Kappa**	**95% confidence Interval for Kappa**	
CC	92.1%	0.76	0.68	0.79	CC	91.7%	0.76	0.70	0.80
EE	90.3%	0.69	0.65	0.70	EE	88.4%	0.70	0.64	0.72
GG	90.7%	0.71	0.67	0.75	GG	91.0%	0.76	0.74	0.80
LI	71.4%	0.28	0.26	0.35	LI	71.1%	0.29	0.24	0.35
SS	77.3%	0.37	0.31	0.43	SS	86.3%	0.61	0.58	0.68
RM	90.2%	0.68	0.63	0.71	RM	89.5%	0.72	0.62	0.75

Abnormal Pigmentation	**Agreement**	**Kappa**	**95% confidence Interval for Kappa**		**Global Assessment**	**Agreement**	**Kappa**	**95% confidence Interval for Kappa**	
CC	92.2%	0.76	0.74	0.79	CC	92.4%	0.77	0.72	0.78
EE	90.0%	0.71	0.68	0.76	EE	91.2%	0.75	0.69	0.76
GG	91.7%	0.76	0.73	0.79	GG	91.9%	0.76	0.74	0.80
LI	81.0%	0.47	0.41	0.55	LI	81.8%	0.50	0.44	0.54
SS	88.6%	0.66	0.63	0.69	SS	87.1%	0.64	0.55	0.69
RM	89.3%	0.70	0.63	0.75	RM	90.2%	0.70	0.69	0.75