British Journal of Dermatology (1994) **130**, 167–173.

Evaluation of cutaneous photodamage using a photographic scale

C.LARNIER, J.-P.ORTONNE,* A.VENOT,† B.FAIVRE,‡ J.-C.BÉANI,§ P.THOMAS, T.C.BROWN¶ AND E.SENDAGORTA**

Produits Roche, Neuilly, France
**Service de Dermatologie, Hôpital Pasteur, Nice, France*
†ECLIMED, Hôpital Cochin, Paris, France
‡Service de Dermatologie, Hôpital Saint-Jacques, Besançon, France
§Service de Dermatologie, Hôpital Michallon, Grenoble, France
¶F. Hoffmann-La Roche Ltd, Basel, Switzerland
***Centre International de Recherche Clinique Roche, Lingolsheim, France*

Accepted for publication 20 July 1993

Summary Clinical assessments of photodamage are based upon a subjective evaluation of characteristic features such as wrinkling and pigmentary change, and are influenced by inter-observer differences in grading criteria. In an effort to standardize the grading of photodamage severity, we have developed a six-point photographic scale in which each of the six grades of overall photodamage severity is depicted by three photographs. The use of three photographs to portray each grade illustrates the diversity and range of manifestations within each grade. This photographic scale was tested by two groups of dermatologists, who used it on two occasions to grade the overall photodamage severity of a single group of female Caucasian subjects. Results indicate high inter-observer agreement, with chance-corrected agreement ranging from 0·44 to 0·63 and from 0·54 to 0·76 on the first and second occasions, respectively. Intra-observer repeatability was high, with chance-corrected agreement ranging from 0·56 to 0·78. Inter- and intra-observer differences were within one category in nearly all cases. Similar grades were assigned by dermatologists with and without experience in treating photodamaged patients. We conclude that application of this scale results in consistent and reproducible clinical evaluations of overall photodamage severity in Caucasian subjects. The scale may be useful in categorizing subjects for epidemiological studies, or in selecting patients for clinical trials.

Clinical assessments of photodamage are based upon a visual and tactile inspection of the skin, and an evaluation of characteristic features such as fine and coarse wrinkling, pigmented lesions, colour, roughness, and telangiectasia.[1–3] Usually, these parameters are integrated into a single score expressing overall severity. Criteria for such evaluations are difficult to quantify on objective scales. Consequently, evaluations made by different clinicians or at different times are influenced by inter- and intra-observer differences, both in the use of grading criteria, and in the way that the grades for each parameter of photodamage (such as wrinkling and pigmentary change) are integrated into a single overall assessment.[4–6]

The consistency, precision, and reproducibility of clinical assessments of photodamage would benefit from the development and conscientious application of well-defined and sensitive grading criteria. Griffiths *et al.*[5] have recently developed a nine-point scale of photodamage severity, illustrated by standardized photographs of five patients depicting five of the nine grades. This photonumeric scale was superior to a written descriptive scale when used to grade photographs of photodamaged individuals, with significant improvements in inter-observer agreement and repeatability. However, even this improved degree of inter-observer agreement and repeatability was not high,[4,7] indicating persistent subjectivity in the way individual graders interpreted and applied visually defined criteria. One probable source of inconsistency is the way in which individual parameters of photodamage contribute unequally and variably to overall severity.[4,8,9] For example, one individual may have extensive wrinkling but little pigmentary change, whereas another may have mild wrinkling but severe mottled pigmentation. Given this variability in the effect of separate photodamage parameters, it is unclear

Correspondence: Dr Catherine Larnier, Produits Roche, 52 boulevard du Parc, 92521 Neuilly-sur-Seine Cedex, France.


1
2
3
1
2
3
1
2
3

Figure 1. Photographic scale used to grade the overall severity of facial photodamage. Three photographs depict each of six grades of photodamage, where grade 1 is mild; grade 2 is mild/moderate; grade 3 is moderate; grade 4 is moderate/severe; grade 5 is severe; and grade 6 is very severe photodamage (Copyright F.Hoffmann-La Roche Ltd).

whether a single reference photograph can be used to define a given grade of photodamage severity; this limitation is exacerbated in the nine-point scale of Griffiths *et al.*, where a single photograph is intended to define more than one severity grade.[5]

We have developed and tested a new six-point photographic scale of overall photodamage severity, in which each of the six grades is depicted by standardized photographs of three representative photodamaged patients. The variable nature of photodamage within each grade is illustrated by the inclusion of a photograph where wrinkling is the primary contributing factor to overall appearance, and a photograph where other factors are more prominent. In addition, age differences between categories are compressed by including photographs of older patients in the milder categories, and photographs of younger patients in the more severe categories. Two groups of qualified dermatologists tested this photographic scale for consistency and reproducibility by using it on two separate occasions to grade the overall photodamage severity of a single group of female Caucasian subjects. Results indicate high inter-observer agreement and intra-observer repeatability.

Methods

Construction of a photographic scale

A total of 988 patients were enrolled in two clinical trials designed to test isotretinoin cream in the treatment of photodamaged skin. Patients were diagnosed as having mild to moderate photodamage (776 patients; 709 females and 67 males),[6,10] or moderate to severe photodamage (212 patients; 192 females and 20 males). Standardized facial photographs of these patients taken at baseline comprised the pool of photographs used to construct this scale. Photographs were taken at each study centre as previously described.[6] Briefly, important features of the photographic procedure included use of a purpose-made stereotactic device with a standard lighting arrangement and camera mount, extensive and explicit instructions regarding the suppression of facial expression, and removal of cosmetics, jewellery, and other extraneous images, training of the study centre staff in the use of the photographic equipment, and prompt review of each developed photograph to ensure high technical quality. All patients had signed a photographic consent form approved by the Institutional Review Boards of the corresponding study centres. From this pool, members of the clinical dermatology staff at F.Hoffmann-La Roche selected photographs of 140 Caucasian women (45° angled view) on the basis of their technical quality and inclusive representation of all grades of photodamage severity.

Four dermatologists, each experienced in treating photodamaged patients, met as a panel and agreed to use a six-point descriptive scale to rate the overall severity of photodamaged skin: mild, mild/moderate, moderate, moderate/severe, severe, and very severe. After discussing the general criteria for inclusion into each category, each dermatologist independently viewed the preselected 140 slides, and assigned grades according to the agreed scale. All four dermatologists then met again as a panel, discussed each photograph and all scores as a group, and assigned a consensus score. Three photographs with appropriate consensus scores were then chosen to depict each of the six grades. Photographs were selected with three goals in mind: (i) each group of three photographs should be clearly distinct from its neighbouring categories in its overall impression of severity; (ii) as far as possible, each group should vary in the severity of individual parameters of photodamage (such as wrinkling and pigmentation); (iii) each series should have a wide age range. Finally, a poster was constructed to contain these 18 reference photographs (Fig. 1).

Testing the photographic scale

Two groups of dermatologists tested the six-point photographic scale for agreement and reproducibility by using it on two separate occasions to grade the overall photodamage severity of a single group of female Caucasian subjects. Subjects rather than photographs were graded, in order to approximate the use of the scale in clinical practice. The first group of graders consisted of three of the four dermatologists, experienced in treating photodamaged patients, who had constructed the photographic scale; these three graders functioned as a panel and assigned a consensus grade to each subject. The second group of graders consisted of eight dermatologists who were not experienced in treating patients with photodamaged skin; these eight graders functioned independently. In this trial, agreement meant that the same grade was assigned by the consensus panel and an independent grader; near agreement meant that the panel and individual grader differed by one category. Repeatability meant that the consensus panel or each independent grader assigned the same grade to a given subject at both assessments; near repeatability meant that the two scores differed by one category.

Sixty female subjects with mild to severe facial photodamage were selected by a dermatologist who did

not participate in the subsequent grading. The six-point photographic guide was used in selection to ensure that subjects represented all grades of severity in approximately equal numbers. These 60 subjects agreed to present for examination on two occasions 8 days apart, and to avoid sun exposure between sessions. Ten days before the first examination, the eight independent graders met for a training exercise in the use of the photographic scale. During the exercise, each grader independently viewed and graded 65 photographs in two sessions, and each session was followed by a group discussion of the assigned grades.

Subjects without make-up or jewellery were examined under standard lighting conditions, and in the same randomized order, by each of the 11 graders in both groups. After examining each subject, the three graders on the panel conferred and assigned a consensus grade; the eight independent graders did not divulge or discuss their grades. The same procedure was followed for the second examination 8 days later, but with the subject order rerandomized. The results of the first examination were not revealed until the second examination was completed.

Inter-observer agreement and intra-observer repeatability were analysed and quantified by use of the kappa coefficient,[11,12] a chance-corrected intraclass correlation coefficient with possible values ranging from -1 (complete disagreement) to $+1$ (complete agreement). Values above 0·75 are generally interpreted as indicating excellent agreement, values between 0·4 and 0·75 indicate fair to good agreement, and values below 0·4 represent poor agreement.[7]

The statistical significance of inter-observer agreement was tested by use of the G statistic defined by Light.[13] The G statistic compares the agreement between each of the individual graders versus the consensus panel, in a way that resembles the calculation of the kappa value, and statistically evaluates the observed agreement for all graders vs. the null hypothesis of random agreement.

Results

Inter-observer agreement for all subjects at assessment 1

During the first assessment, one of the 60 subjects was not examined by the three graders on the consensus panel, and was therefore excluded from the analysis. Table 1 presents the distribution of consensus grades assigned by the panel at the first assessment; all grades were represented, although grades 5 and 6 (severe and very severe) contained the fewest subjects.

Table 1. Assessment 1. Consensus panel evaluation of photodamage severity for all 59 evaluated subjects

Photodamage category	*n* (%)
1	10 (16·9)
2	11 (18·6)
3	14 (23·7)
4	14 (23·7)
5	5 (8·5)
6	5 (8·5)
Total	59 (100)

Table 2 presents a comparison of grades assigned by the consensus panel vs. each of the eight individual graders. Agreement between individual graders and the consensus panel was scored for 33–42 of the 59 evaluated subjects, corresponding to kappa values ranging from 0·46 to 0·64 (median 0·59). These results indicate fair to good agreement in all cases. Near agreement (within one category) was scored for all but two cases (data not shown). Analysis of the pooled results for inter-observer agreement at assessment 1 yielded a highly significant G value of 23·6 ($P<0·001$).

Inter-observer agreement for 51 subjects graded at both assessments

Fifty-one subjects returned for the second examination 8 days later. Subjects who did not return had been classified as grade 1 (three subjects), grade 2 (four subjects), and grade 4 (one subject) by the consensus panel at the first assessment. Because these eight subjects were unequally distributed among categories, inter-observer agreement at assessment 1 was recalculated for the 51 returning subjects. Table 3 summarizes

Table 2. Assessment 1. Inter-observer agreement for all 59 evaluated subjects

Grader	p_o/N_T*	Kappa
1	40/59	0·603
2	36/59	0·518
3	39/59	0·582
4	41/59	0·622
5	41/59	0·622
6	33/59	0·455
7	37/59	0·537
8	42/59	0·642

* Number of scores in agreement with the consensus panel (p_o) divided by total number graded (N_T).

Table 3. Assessments 1 and 2. Inter-observer agreement for all 51 subjects evaluated at both assessments

	Assessment 1		Assessment 2	
Grader	p_o/N_T*	Kappa	p_o/N_T*	Kappa
1	34/51	0·586	32/51	0·544
2	29/51	0·468	33/51	0·561
3	33/51	0·567	41/51	0·757
4	34/51	0·583	33/51	0·562
5	34/51	0·585	37/51	0·658
6	28/51	0·442	38/51	0·689
7	31/51	0·512	33/51	0·565
8	36/51	0·633	39/51	0·708

* Number of scores in agreement with the consensus panel (p_o) divided by total number graded (N_T).

the inter-observer agreement at assessment 1 and assessment 2 for the 51 subjects who were graded on both occasions.

At assessment 1, agreement between individual graders and the consensus panel was scored for 28–36 of the 51 evaluated subjects, corresponding to kappa values ranging from 0·44 to 0·63 (median 0·58). Analysis of the pooled results yielded a highly significant G value of 23·2 ($P<0·001$). These results are similar to those presented in Table 2 for all 59 subjects graded at that assessment.

At assessment 2, agreement between individual graders and the consensus panel was scored for 32–41 of the 51 evaluated subjects, corresponding to kappa values ranging from 0·54 to 0·76 (median 0·61). These results indicate good agreement in all cases. Near agreement (within one category) was scored for all but one case (data not shown). Analysis of the pooled results for inter-observer agreement at assessment 2 yielded a highly significant G value of 26·8 ($P<0·001$).

Table 4. Assessments 1 and 2. Inter-observer agreement for 39 subjects with identical consensus panel scores at both assessments

	Assessment 1		Assessment 2	
Grader	p_o/N_T*	Kappa	p_o/N_T*	Kappa
1	28/39	0·658	27/39	0·627
2	23/39	0·501	25/39	0·560
3	29/39	0·687	32/39	0·780
4	27/39	0·625	28/39	0·662
5	29/39	0·687	31/39	0·750
6	23/39	0·502	33/39	0·814
7	24/39	0·533	28/39	0·654
8	29/39	0·687	31/39	0·748

* Number of scores in agreement with the consensus panel (p_o) divided by total number graded (N_T).

Table 5. Intra-observer repeatability

Grader	p_o/N_T*	Kappa
Panel	39/51	0·708
1	37/51	0·663
2	35/51	0·604
3	41/51	0·758
4	36/51	0·628
5	42/51	0·779
6	33/51	0·564
7	35/51	0·604
8	36/51	0·631

* Number of subjects assigned the same grade at both assessments (p_o) divided by total number evaluated at both assessments (N_T).

Inter-observer agreement for a reference subpopulation of 39 subjects

Thirty-nine of the 51 subjects (76%) who presented for both evaluations were assigned the same grade on both occasions by the consensus panel. Inter-observer agreement at assessments 1 and 2 was calculated for these 39 subjects to test the hypothesis that this subpopulation was particularly amenable to consistent grading, using the photographic scale, perhaps due to the presence of clear distinguishing signs of photodamage severity. Table 4 summarizes inter-observer agreement at both assessments for this subpopulation; results do not appear to be substantially different from those obtained for the larger cohort of 51 evaluated subjects (Table 3).

Intra-observer repeatability

Table 5 presents a comparison of grades assigned to each subject at the two assessments. Intra-observer repeatability is calculated for the consensus panel and for each of the eight individual graders. Grades assigned by the consensus panel were repeatable in 39 of 51 cases, corresponding to a kappa value of 0·71. For the eight individual graders, consistent assignments were made in 33–42 of the 51 evaluated cases, corresponding to kappa values of 0·56–0·78 (median 0·63). These results indicate good intra-observer repeatability in all cases. Near repeatability (within one category) was scored in all cases for the consensus panel and in all but one case for the eight individual graders (data not shown).

Discussion

Results of this study indicate that application of a newly devised, six-point photographic scale results in consistent and reproducible clinical evaluations of overall photodamage severity. We believe that the advantage of this new scale over previous photographic guides rests in the use of three photographs, rather than one, to depict each degree of severity. Because grading overall severity requires a composite evaluation of independently variable elements (i.e. wrinkles, pigmentation), comparing a patient with multiple reference photographs facilitates consistent weighting and integration of these factors in estimating an overall score.

Grading repeatability was approximately equal for the panel of graders experienced in treating photodamaged patients vs. individual graders with no particular experience. This result indicates that the six-point scale is robust, and can be applied by clinicians without specialized expertise. Agreement between the consensus panel and individual graders was typically within one category, indicating that evaluations made by a single grader do not require subsequent review and confirmation.

We have not used our photographic scale to evaluate treatment response for patients receiving therapy for photodamaged skin. Instead, we have assessed treatment response by use of within-patient comparisons, in which each patient's baseline condition acts as a comparative control for subsequent improvement. Side-by-side comparisons of standardized photographs taken before and after treatment, and rated by a panel of experienced graders[6] has proved to be a particularly useful and reliable measure of treatment efficacy, especially when combined with other within-patient assessments such as the physician's evaluation of response, using baseline photographs for comparison, and the patient's self-assessment.[10] Although we have not directly compared the sensitivity of our six-point categorical scale with that of within-patient evaluations, we believe that within-patient comparisons can detect either subtle or marked changes, whereas the sensitivity of the categorical scale is inherently limited to the detection of improvement equal to or greater than the difference between categories.

Although the photographic scale has only been tested with female Caucasian subjects, we see no reason why it cannot be applied to grade photodamage severity in male Caucasians, as the principal manifestations of photodamage (wrinkling, pigmentary change, roughness) occur in both genders. Minor systematic differences between males and females in the effects of chronological ageing and in the relationship between sun exposure and the development of clinical signs of photodamage should affect grading in a uniform rather than an erratic way, and this would not diminish the consistency and reproducibility of ratings. As discussed by Goh[9] and by Griffiths *et al.*,[5] the principal manifestation of photodamage in Far East Asians is pigmentary change rather than wrinkling, and therefore our photographic scale should probably not be used to grade photodamage in Asian patients. Instead, separate photographic scales may be necessary for non-white Caucasian populations.

The six-point photographic scale should be useful in categorizing white Caucasian subjects in epidemiological studies where photodamage severity is either studied or thought to be a factor in the relevant endpoint. The scale will also be useful in characterizing white Caucasian populations entering clinical trials for treatment of photodamage, or for conditions where photodamage severity is a relevant cofactor.

Acknowledgments

Jean-Claude Peyrieux (Statmed, Lyon, France) and Roy Ward (Roche International Clinical Research, Lingolsheim, France) provided invaluable assistance and advice in data processing and statistical analysis.

References

1 Gilchrest BA. Skin aging and photoaging: an overview. *J Am Acad Dermatol* 1989; **21:** 610–13.
2 Leyden JJ. Clinical features of ageing skin. *Br J Dermatol* 1990; **122** (Suppl. 35): 1–3.
3 Taylor CR, Stern RS, Leyden JJ, Gilchrest BA. Photoaging/photodamage and photoprotection. *J Am Acad Dermatol* 1990; **22:** 1–15.
4 Stern RS, Coopman SA. The measure of youth. *Arch Dermatol* 1992; **128:** 390–3.
5 Griffiths CEM, Wang TS, Hamilton TA *et al.* A photonumeric scale for the assessment of cutaneous photodamage. *Arch Dermatol* 1992; **128:** 347–51.
6 Armstrong RB, Lesiewicz J, Harvey G *et al.* Clinical panel assessment of photodamaged skin treated with isotretinoin using photographs. *Arch Dermatol* 1992; **128:** 352–6.
7 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33:** 159–74.
8 Gilchrest BA. The variable face of photoaging: influence of skin type. *Cosmet Toilet* 1992; **107:** 41–2.
9 Goh SH. The treatment of visible signs of senescence: the Asian experience. *Br J Dermatol* 1990; **22** (Suppl. 35): 105–9.
10 Sendagorta E, Lesiewicz J, Armstrong RB. Topical isotretinoin for photodamaged skin. *J Am Acad Dermatol* 1992; **27:** S15–18.
11 Cohen J. A coefficient of agreement for nominal scales. *Educ Physiol Measurement* 1960; **20:** 37–46.
12 Fleiss JD, Cohen J. Large sample standard errors of kappa and weighted kappa. *Psychol Bull* 1969; **72:** 323–7.
13 Light RJ. Measures of response agreement for qualitative data: some generalizations and alternatives. *Psychol Bull* 1971; **76:** 365–77.